Surveillance Video Summarization based on Trajectory Rarity Measure

Gerar Francis Quispe Torres

Advisor: MSc. Rensso Victor Hugo Mora Colque

Committee Members:
Ph.D. Javier Montoya Zegarra – ETH Zurich – Suiza
Ph.D. Guillermo Cámara Chávez – Universidade Federal de Ouro Preto – Brazil
Ph.D. Juan Carlos Gutiérrez Cáceres – Universidad Católica San Pablo – Peru
Ph.D. José Eduardo Ochoa Luna – Universidad Católica San Pablo – Peru

Thesis submitted to the
Departament of Computer Science
in partial fulfillment of the requirements for the degree of
Master in Computer Science.

Universidad Católica San Pablo – UCSP
April of 2019 – Arequipa – Peru

This thesis is dedicated to my family, to Computer Science masters program, Universidad Católica San Pablo, Arequipa, Perú and Smart Surveillance Interest Group. Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil.

Acronyms

AP Affinity Propagation

DFT Discrete Fourier Transform

DTFT Discrete-time Fourier Transform

DWT Discrete Wavelet Transform

GPS Global Positioning System

HOFM Histograms of Optical Flow Orientation and Magnitude

HOG Histogram of Oriented Gradients

HOOF Histogram of Oriented Optical Flow

IDFT Inverse Discrete Fourier Transform

KDE Kernel Density Estimation

PLSIC Partial Least Squares Image Clustering

RGB Red, Green, Blue

SHNN-CAD Sequential Hausdorff Nearest-Neighbor Conformal Anomaly Detector

SIFT Scale-Invariant Feature Transform

SOM Self Organized Map

SVM Support Vector Machine

t-SNE t-Distributed Stochastic Neighbor Embedding

Acknowledgements

First of all, I want to thank God for guiding me through these three years of study.

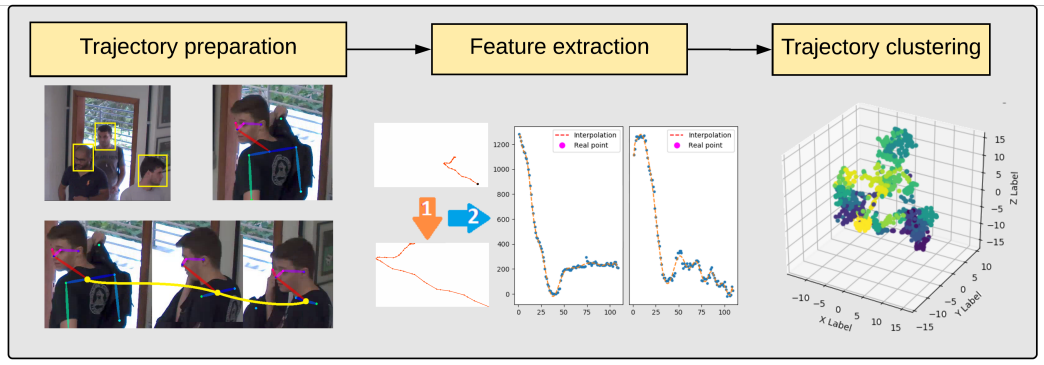
I would like to thank in a special way to the National Council for Science, Technology and Technological Innovation (CONCYTEC-PERU) and to the National Fund for Scientific Development, Technological and Technological Innovation (FONDECYT-CIENCIACTIVA), which through the Management Agreement 234-2015-FONDECYT have allowed the grant and financing of my studies in the Master Program in Computer Science at Universidad Católica San Pablo (UCSP).

I would like to express my sincere gratitude to my advisor Rensso Mora and my co-advisor Prof. William Schwartz for the continuous support during my Master's degree studies and research; his patience, enthusiasm and comprehension were fundamental for me to get this work done.

I would also like to thank all the friends I met during my stay in Belo Horizonte (Brasil) and Arequipa (Perú), we shared lot of invaluable academic, professional, and social experiences. I also thank my friends from undergrad school.

Last but not the least, I would like to thank to my family for providing me with unfailing support and continuous encouragement.

Abstract



The dynamic video summarization of surveillance videos has several critical applications, mainly due to the wide availability of digital cameras in environments such as airports, train and bus stations, shopping centers, stadiums, buildings, schools, hospitals, roads, among others. This study presents an approach for the generation of dynamic summary on surveillance video domain based on human trajectories. It has an emphasis on trajectory descriptors in conjunction with the unsupervised clustering method. Our approach contribute to existing literature concerning the combination of methods and objectives. We hypothesize that the clustering of trajectories permits to identify rare trajectories base on their morphology. The clustering as an output provides numerous subsets of trajectories or clusters and the number of elements of a specific cluster is used to determine their rarity. Those subsets with few components are rare while the others that have a high number of elements are considered ordinary; therefore, the implications of our study show that is possible to use unsupervised clustering for automatic detection of rare trajectories based on their morphology and with this information segment videos. We experimented with different sets of trajectories segmenting the rare videos from our ground truth.

Keywords: Dynamic surveillance video summarization, Trajectory clustering, Morphology trajectory descriptor, Trajectory feature extraction.

Resumen

El resumen dinámico de vídeos de vigilancia tiene múltiples aplicaciones importantes, principalmente debido a la amplia disponibilidad de cámaras digitales en entornos como aeropuertos, estaciones de ómnibus, estaciones de trenes, centros comerciales, estadios, escuelas, hospitales, autopistas entre muchos otros. Este estudio presenta un abordaje para la generación de resúmenes dinámicos de vídeos en un dominio de vigilancia utilizando trayectorias. Enfatizamos nuestro estudio en la búsqueda de características que sean adecuadas para que conjuntamente con los métodos de agrupación no supervisados nos permitan segmentar información. Nuestro abordaje contribuye a la literatura considerando la combinación de métodos y objetivos. Utilizamos el agrupamiento automático de trayectorias para generar subconjuntos e identificar aquellos que son raros, este agrupamiento está basado en la morfología de cada trayectoria, el número de elementos de un subconjunto nos servirá para determinar su raridad. Los subconjuntos que contienen un número diminuto de elementos son considerados raros mientras que los restantes, que son aquellos que poseen un número mayor de elementos, son considerados ordinarios. Las implicaciones de nuestro estudio muestran que es posible utilizar métodos de agrupamiento para la detección automática de trayectorias raras. La información de número de elementos de cada grupo será utilizada para reconocer vídeos raros; además, las trayectorias utilizadas en nuestros experimentos no tienen ningún demarcado previo. La eficacia de los enfoques presentados se muestran a través de experimentos y observaciones en distintos conjuntos de trayectorias que evidencia de la utilidad de nuestro aporte.

Palabras clave: Resumen dinámico de vídeos de vigilancia, agrupación de trayectorias, descriptor de morfología de trayectorias, extracción de características en trayectorias.

Contents

List of Tables

List of Figures

Chapter 1

Introduction

Nowadays, the rapid development of digital video and the need for faster browsing of a large amount of data and content indexing has led to the need for effective and advanced techniques for analysis and video retrieval. Since multimedia sources are growing, the delivery of content and video retrieval has become very slow. The significant technological advances increased the use of videos and the human effort taken to process it. This problem has promoted a generation of new trends and innovations in summarization methods. At the same time, the treatment of surveillance videos is significant today because it is one effective tool to help maintain the security in general. The worldwide market for video surveillance camera equipment grew by 7% in 2016 (IHS, 2016). These cameras are placed in bars, banks, casinos, schools, hotels, airports, hospitals, restaurants, military installations and commercial stores to name a few (Torres *et al.*, 2016).

Many times the surveillance operators have to look for specific events either because a crime has happened or there is a criminal investigation; sometimes they are in search of some rare event without the certainty of finding it. This task is time-consuming because the videos have to be viewed sequentially one by one wasting time watching repeated scenes within not relevant content. Some limitations of using the surveillance cameras are the crime prevention effect (Welsh & Farrington, 2008) and the deterrence of terrorists (Stutzer & Zehnder, 2010), since we process events that happened or are happening but not a model that predicts crimes. The cameras detect every event throughout the day or after the event has occurred; on the other hand, the suspects spoofing their identity. They try not to collaborate with the surveillance video systems. Despite all of this, surveillance videos demonstrated are suitable for forensic purposes.

The surveillance video operators have to maintain a level of concentration and divide that attention to monitor for multiple types of occurrences at a single location; hence, humans have limitations due to fatigue at monitoring for rare events across multiple video streams and the error rate produced fluctuates depending on the different unpredictable circumstances. In 1999, Chabris & Simons (2011) exhibited a phenomenon called "*inattentional blindness*" this study explain that when the atten-

tion of someone is on an object or task, they often fail to perceive an unexpected object even if it is in the middle of their field of sight. In this study work, observers were watch a video of recording of people passing a basketball amongst themselves. They were asked to count the number of passes. During the video, someone in a gorilla suit walks into the middle of the group and beats its chest. Less than half of the observers noticed that there was a gorilla. Theses studies have shown that humans have some deficiencies that make them poorly suited to being multiple cameras operators. These deficiencies get worse as the operators monitor many cameras at the same time and in some cases, the operator wastes time watching irrelevant content. The poor configuration of technology, the significant amount of data, the lack of system integration, the lack of operators and the growth of cameras (Keval & Sasse, 2010) are considerable factors that make these tasks difficult and exhausting. Wallace *et al.* (1998) demonstrated that when a human is operating sixteen cameras, they cannot operate more than thirty minutes efficiently, after this a break is necessary for health reasons. Dee & Velastin (2008) determined that the number of cameras per operator vary from four until seventy-eight, demonstrating that the performance of the operator increases with a smaller number of cameras. An operator can monitor from one to four cameras at the same time without having much difficulty. The automatic recovery of videos becomes essential at this point since it serves as an assistant and helps with the proper handling of information. The automatic recovery of videos based on semantic information is an area that is yet to be explored deeply.

Various automatic and intelligent methods have been studied and developed to aid surveillance camera operators as human activity recognition (Aggarwal & Ryoo, 2011), detection of anomalous events (Sodemann *et al.*, 2012), gesture recognition (Tung & Ngoc, 2014), movement analysis (Fortun *et al.*, 2015), re-identification of people (Bedagkar-Gala & Shah, 2014), analysis of urban traffic (Tan *et al.*, 2016), background subtraction in videos (Sobral & Vacavant, 2014), object tracking (Wu *et al.*, 2013) among others. For instance, on a specific topic in the treatment of surveillance videos, there is the detection of anomaly trajectories (Piciarelli *et al.*, 2008; Laxhammar & Falkman, 2014; Ergezer & Leblebicioğlu, 2016; Sillito & Fisher, 2008). In the future it is expected that the deeper analysis and the creation of new methods in the field of computer vision and science in general can solve the problem that surveillance video has, it in real life or at least in a controlled environment.

Video summarization has become a tool to surveillance systems in helping camera operators (Murugan *et al.*, 2018; Ajmal *et al.*, 2017). Video summarization technique is an alternative to reduce power consumption, human effort and time consumption since video viewing devices consume energy and camera operators spend time and effort to find rare events. They have to look for information within a much larger set of videos demanding concentration. Video summarization is used in others fields of study like the detection of anomalies, classification of human actions or understanding activities (Turchini *et al.*, 2015). The digital video revolution has grown enormously in recent years creating applications and new research areas that focus on effectiveness and efficiency. This is due to the considerable computer power required to create video summaries. During the last few years, the amount of video content in multimedia systems has increased dramatically, as well as a large number of commercial sources. These sources

have diversity and heterogeneity in their information making difficult the search and retrieval of specific content. People become a potential content sharing contributor through different communication platforms (Eppler & Mengis, 2004). Nowadays we can find video content in digital libraries, personal collections, social networks, websites, optical storage discs like DVD, Bluray, digital television and many others. The steadily trend of grown will continue in the coming years. Consequently, we are overwhelmed with an enormous and increasing amount of video information (Rogers *et al.*, 2013). A video is by nature time-consuming media (Austin *et al.*, 2016), this task often makes data management difficult. For that reason, video summaries seem to be adequate for efficient access and navigation of data allowing the users to recover the content efficiently.

The literature presents researchers that approached the problem of video summarization (Sebastian & Puthiyidam, 2015); however, there are few studies on automatic video summarization oriented to surveillance domain. Many of the surveillance videos are always recording, and this trend is growing. Recorded sequences can become long and tedious to review; this is one crucial reason to gather data automatically. It demands a correct and substantial data representation also called semantic information extraction. These tools significantly improve the surveillance work since it is clear that when the operators have a summary and substantial information of a specific event, they make decisions with greater certainty and velocity. These methods are still an open and challenging task in literature.

The summarization in surveillance videos is different from the summarization in other domains. An sample of this is that surveillance videos have few background changes especially for non-static cameras. In summarization of cinematographic video, it is essential to take into account the background, the changes of colors, the scenes of the video, among other factors. These characteristics do not happen in the surveillance domain. In surveillance domain is possible to guide the summarization in to cut scenes with semantic information, and these make it using characteristics that depend on their content either by object detection, behavior recognition, face recognition, object classification or object tracking among others (Murugan *et al.*, 2018). As our hypothesis is different, we must perform other processes, introducing further level information such as recognition of events or representation of trajectory information extracted. It is also necessary to note that this process of summarization is beneficial to operators and our model contributes in this direction. The rareness is defined as something that occurs at a lower frequency and therefore of our interest.

In this context, we propose a model based on human trajectories. The idea is to describe trajectories by their morphology using this information to cluster the trajectories. A significant contribution of our proposed approach is to segment videos according to the human trajectories in a video; hence, our model retrieves a group of videos based on the presence of a person in the sequence frames. This process may aid operators to differentiate the videos, for instance, people running, loitering or unusual event as people walking in different directions. Grouping trajectories in our model allows us to determine strange behavior. The descriptors provide feature vectors from trajectories. The model can separate the events that have a higher degree of rarity from

the repeated videos. In order to segment rare trajectories, the approach employs clustering methods to group them. The model allows classification of trajectories to detect the uncommon, this process uses unsupervised techniques. Clustering can distinguish the different subgroups of characteristics extracted from the surveillance videos that belong to our database.

Our hypothesis consists of the following proposal: *The clustering of trajectories permits the identification of rare trajectories in a dataset, in consequence the clustering provides numerous subsets of trajectories and depending on the number of elements that these subsets have, it is possible to classify how uncommon they are. Those subsets with few numbers of elements represent infrequent events while the others who have a high number of elements will be considered ordinary.* The descriptor should be sensitive to model trajectory morphology and this type of summarization should be based on semantic information.

Note that video summarization in surveillance videos is different for classic models of summarization. Indeed summarization takes into account the change of the scenes; however, in surveillance video, the background does not change on static camera and operators are always observing people. Therefore, the summarization for surveillance video can be seen from another perspective. Hours of surveillance video can be shifted to focus on a single event; for example, moments in which people are running or moments when people are waiting for a long time in the same place.

Finally, we analyzed qualitative and quantitative results in the evaluation chapter, addressing the problem of dynamic summarization for surveillance video using SSIG-dataset and our ground truth having considerable results.

1.1 The problem

There is more than one problem when someone wants to achieve a summary of surveillance video. One of them is the problem of determining a video feature that allows the reduction of search data. This problem is referred to as extracting semantic content that can be used to highlight or disregard video segments. For that purpose, our model employs the semantic information of trajectories that aggregates meaning for summarization. This task involves finding a correct combination of the descriptor and clustering method. All of this happens in the characteristic space denominated hyper-plane in feature engineering field.

There are two problems with trajectory descriptions, the collection and data representation. The collection consists on describe a person movement with trajectories, in this part is important not confuse points since the extraction of points with detection can be confused. This extraction is from the videos. At the end of the collection the detected points of a trajectory should belong to the same person. On the other hand, the second problem is to describe or to represent the trajectories in order to classify them by their morphology. In the literature there are models that create trajectory

representations in many forms; nevertheless, none of these representations takes a trajectory like a unit; it means, our model does not take into account the location of trajectory points but rather the form that this trajectory acquires.

The location information of trajectory points can limit the description of the morphology; for example, the trajectories themselves recover the position of the object over time and contain significant information for the cluster; however, directly performing clustering on those positions provides lower results. According to Xu *et al.* (2015), two reasons lead to that simple conclusion: a) *Variation within the cluster*, due to the uncertain nature of the movement of the object and the vagueness of the trajectory extraction, trajectories of the same clusters can have high variation. b) *Ambiguity between clusters*, in many scenarios, trajectories with related shapes and near positions may belong to different clusters due to its underlying semantic differences. Video summarization is a complex problem that requires specific solutions according to each domain. The approach depends on the video content. This is a continuing problem since the meanings change depending on the contexts and the semantic levels of information are difficult to extract. In video summarization, the concept of the essential parts of a video changes according to the viewer, within the literature found there is no exact metric to measure the quality of an automatic summary.

1.2 Justification and motivation

We already saw that the processing of surveillance videos is required and it is one valuable tool to help in monitoring the security on the streets. The camera operators have to look for specific events either because a crime has happened or they are doing a crime investigation. When referencing Xu *et al.* (2015), we verify the few existing studies on this topic in scholarly articles considering the treat of trajectories as an indivisible unit and contemplate any alterations on their morphology. Our approach describes a whole trajectory considering this aspects, this characteristics are not common treated on literature. We consider that it is appropriate to pre-process the trajectories in order to highlight their characteristics, this can be contrasted with Xu *et al.* (2015) : "*Although the trajectories themselves recover the position of the object over time and contain significant information for the cluster, directly performing clustering on those positions provide poor results*" .

Our approach addresses the problems related to information retrieval since we obtain information that is more substantial in videos. To achieve this task, our approach uses semantic information. We call semantic information to the information extracted from the content of the videos, in this case we refer to people walking in the videos generating different trajectories. This information permits us to classify and identify the rarity in a real space of characteristics. For instance, others approaches take the Red, Green, Blue (RGB) information as features that are used to perform their summarizations. Represent semantic information is a challenging problem; it is latent and complicated since it could have different meanings according to the context. Our research focuses on these properties.

In traditional surveillance systems, the review of surveillance videos is done based on an operator who manually reviews the videos one by one, which is tedious and error-prone. We store only the scenes where something strange happens and skip the repeated or everyday scenes. This research therefore benefits the camera operators. They will spend less time on information retrieval. After extracting semantic information from our videos, we focus clustering it correctly. Semantic information can be challenging to extract compare with other information but it provide us with valuable information.

An aspect to highlight in this section is the relationship between the detection of the rarity with video summarization. These concepts are related due to the interest of the camera operator in detecting unusual events as part of their work since the rare events contribute with knowledge for operators. In other words, the unusual events are considered important for the camera operators. Different queries are outside of our approach. Strictly speaking, the meaning of anomaly is not the same as the rarity. However, they share some similarities. In the surveillance domain, the rarity is considered to collect knowledge of a video and our proposed approach seeks to extract those significant events for the observer. Video summarization is a challenging problem and is solving for a specific domain as we mention this is a latent problem and is yet explored deeply.

1.3 Objectives

The overall objective of this study is to propose an approach for automatic summarization of surveillance video based on trajectories. The detection of uncommon events are involved in this process, and the trajectories are defined as rare by a clustering method. This permits the segmentation of information based on rare morphology trajectories from our video database. Our model extracts uncommon events shown these with priority within a video summary. These uncommon events are significant and concise to provide video summary in the surveillance domain.

1.3.1 Specific objectives

1. Detect rare trajectories considering them as indivisible units.

2. Investigate trajectory descriptors that can correctly differentiate trajectories by their morphology.

3. Investigate and evaluate different unsupervised clustering techniques with an unknown number of clusters to classify the trajectories and to decide which is the best.

4. Selecting a metric to measure the quality of our clustering methods and the evaluation of our results on a proposed database.

1.4 Text organization

Our text is organized as come next, on Chapter 2 is describing concepts that help us to better understand our study since these concepts will be mentioned as we go through the text reading, we try explain this concepts in a friendly way. Relevant concepts associated with the investigated problem are presented and discussed in Chapter 3, this begins describing the importance of the treatment of large amounts of data and how summarization on surveillance videos is becoming increasingly crucial over time. Subsequently, we will present an introduction to the classification of video summarization methods found in the scholarly articles, then the concepts involved in video summarization in general, related study works and also the studies with the treatment of trajectories are described. Chapter 4 will begin presenting the relation of our methodology with the literature and a description of all methods involved in our model. Experiments are detailed in Chapter 5, including our creation of ground truth, evaluation methodology and results. Finally, Chapter 6 concludes the study with some final remarks and directions for future study works.

Chapter 2

Background

This chapter presents concepts related to our model in order to familiarize the reader with the themes of this research.

2.1 Discrete Fourier Transform

The Fourier transform is a mathematical function that decomposes a waveform, which is a function of time, into the frequencies that make it up. The result produced by the Fourier transform is a complex valued function of frequency. The absolute value of the Fourier transform represents the frequency value present in the original function and its complex argument represents the phase offset of the basic sinusoidal in that frequency. The Fourier transform is also called a generalization of the Fourier series. This term can also be applied to both the frequency domain representation and the mathematical function used. The Fourier transform helps in extending the Fourier series to non-periodic functions, which allows viewing any function as a sum of simple sinusoids. The Fourier transform of a function $f(x)$ is given by:

$$f(x) = \int_{-\infty}^{+\infty} F(k) e^{2\Pi i k x} dk \tag{2.1}$$

$$F(k) = \int_{-\infty}^{+\infty} f(x) e^{-2\Pi i k x} dx \tag{2.2}$$

where $F(k)$ can be obtained using inverse Fourier transform.

In mathematics, the Discrete Fourier Transform (DFT) is a type of discrete transformation used in Fourier analysis. It works by transforming one mathematical function into another, obtaining a representation in the frequency domain where the domain of the original function is the time. The DFT requires that the input function be a discrete

sequence of finite duration. These sequences are usually generated from the sampling of a continuous function, such as the human voice. On the other hand, the Discrete-time Fourier Transform (DTFT) is a transformation that only evaluates sufficient frequency components to reconstruct the finite segment analyzed. DFT is used to analyze a single section of a periodic signal that extends infinitely. If this is not accomplished, one more window could be used to reduce the noise in the spectrum. For the same reason, the inverse of DFT: The Inverse Discrete Fourier Transform (IDFT), cannot reproduce the full-time domain unless the signal entry were indefinitely periodic. For these reasons, the DFT is a Fourier transformer for analysis of discrete time signals in unlimited domain. The base sinusoidal functions that arise from the decomposition have the same properties.

Definition 2.1.1. *Let be $x_0, x_1, ..., x_{n-1}$ complex numbers, the Discrete Fourier Transform is defined as:*

$$X_k = \sum_{n=0}^{N-1} x_n e^{-\frac{2\pi i}{N} nk} \qquad k = 0, ..., N-1$$

.

The DFT is the equivalent of the continuous Fourier Transform for signals known only at N instants separated by sample times (i.e. a finite sequence of data). Let $f(t)$ be the continuous signal which is the source of the data. Let N samples be denoted $f[0], f[1], f[2], ..., f[k], ..., f[N-1]$. The Fourier Transform of the original signal, $f(t)$, would be

$$F(jw) = \int_{-\infty}^{+\infty} f(t) e^{-jwt} dt \tag{2.3}$$

We could regard each sample $f[k]$ as an *impulse* having the area $f[k]$. Therefore the integral exists only at the sample point:

$$F(jw) = \int_{0}^{(N-1)T} f(t) e^{-jwt} dt \tag{2.4}$$

$$= f[0]e^{-j0} + f[1]e^{-wT} + ... + f[k]e^{-jwkT} + ...f(N-1)e^{-jw(N-1)T} \tag{2.5}$$

i.e.

$$F(jw) = \sum_{k=0}^{N-1} f[k] e^{-jwkT} \tag{2.6}$$

We could, in principle, evaluate this for any ω, but with only data points to start with, only final outputs will be significant. You may remember that the continuous Fourier transform could be evaluated over a finite interval (usually the crucial period

T_0) rather than from $-\infty$ to $+\infty$ if the waveform was periodic. Similarly, since there are only a finite number of input data points, the DFT treats the data as if it were periodic (i.e. $f(N)$ to $f(2N-1)$ it is the same as $f(0)$ to $f(N-1)$).

2.2 Discrete Wavelet transform

A wavelet is a mathematical function useful in digital signal processing and image compression. The use of wavelets for these purposes is a recent development, although the theory is not new. The principles are similar to those of Fourier analysis, which was first developed in the early part of the 19th century. In signal processing, wavelets make it possible to recover weak signals from noise . This has proven useful especially in the processing of X-ray and magnetic-resonance images in medical applications. Images processed in this way can be "cleaned up" without blurring or muddling the details. In Internet communications, wavelets have been used to compress images to a greater extent than is generally possible with other methods. In some cases, a wavelet-compressed image can be as small as about 25 percent the size of a similar-quality image. The best way to introduce wavelets is through their comparison to Fourier transforms a common signal analysis tool. Wavelet and Fourier transform represents a signal through a linear combination of their basic functions. Like the Fourier Transform, the Wavelet Transform decomposes signals as a superposition of simple units from which the original signals can be reconstructed. The Fourier Transform decomposes signals into sine and cosine functions of different frequencies, while the Wavelet Transform decomposes signals into wavelets. The wavelet transform base functions are compact, or finite in time, while the Fourier sine and cosine functions are not. This feature allows the Wavelet Transform to obtain time information about a signal in addition to frequency information. Since the Fourier Transform is a global integration transformation and there is no time factor in it, it cannot effectively analyze non-stationary signals whose statistical properties change with time. To analyze non-stationary signals, we need to decompose signals into units that are localized in both time and frequency domains. The Fourier Transform is widely used in science and engineering to analyze and process signals. It is a global integral transformation of the form:

$$X(f) = \int_{-\infty}^{+\infty} x(t) e^{-j2\pi ft} dt \tag{2.7}$$

This decomposes the original signal into sine and cosine signal units of different frequencies. These signal units are more accessible to analyze and process than the original complex signal. The DWT, on the other hand, has a window size that varies frequency scale. This technique is advantageous for the analysis of signals containing both discontinuities and soft components. Short high-frequency base functions are needed for discontinuities, while at the same time, long low-frequency ones are needed for the soft components.

In conclusion, Wavelets are a class of a functions used to localize a given function

in both space and scaling. A family of wavelets can be constructed from a function $\Psi(x)$, sometimes known as a "mother wavelet," which is confined in a finite interval. "Daughter wavelets" $\Psi^{a,b}(x)$ are then formed by translation (b) and contraction (a). Wavelets are especially useful for compressing image data, since a wavelet transform has properties which are in some ways superior to a conventional Fourier transform.

An individual wavelet can be defined by

$$\Psi^{a,b}(x) = |a|^{-1/2}\Psi(\frac{x-b}{a}). \tag{2.8}$$

Then

$$W_\Psi(f)(a,b) = \frac{1}{\sqrt{a}}\int_{-\infty}^{\infty} f(t)\Psi(\frac{t-b}{a})dt, \tag{2.9}$$

and Calderón's formula gives

$$f(x) = C_\Psi \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} < f, \Psi^{a,b} > \Psi^{a,b}(x)a^{-2}dadb. \tag{2.10}$$

A common type of wavelet is defined using Haar functions.

2.3 Self Organized Maps

Self organized maps is a Feed-forward neuronal network (Kohonen, 1998). This technique does not perform feedback or process back-propagation, it is an unsupervised training network which learns to form their classification of the training data without external help. This neuronal network learns the features of each trajectory in an unsupervised form to produce a discrete distribution of the feature space that it receives as input. This distribution is a so called two dimensional quadrille map. The self-organized maps are different from other artificial neural networks in the sense that they use a neighbourhood function to preserve the topological properties of the input space. The technique assigns each neuron a vector of characteristics. The dimension of these characteristics is the same dimension that the input vector of the neuronal network has. The procedure for locating one vector for the input data in the map consists of finding the neuron with the nearest vector of weights to the vector of the input data space; i.e., the smaller metric distance between neurons. Amongst others, neuronal network and self-organized-maps operate in two modes: training and mapping. In training, the map is built using training examples. While in the mapping, it classifies one new entry. Self Organized Map (SOM) assumes that the input patterns sharing common features define the class and the network identifies those features across the range of input patterns. In competitive learning the output neurons compete among

themselves for activation. This is an exciting class of an unsupervised system. At the end of a one-time competition, only one neuron is activated. This activated neuron is called a "winner-takes-all" neuron or just the "winning" neuron. Such competition can be induced or implemented by having lateral inhibition connections (negative feedback paths) between the neurons. The result is that the neurons are forced to organize themselves. For obvious reasons, such a network is called a SOM.

The principal goal of a SOM is to transform an incoming signal pattern of arbitrary dimension into a one or two-dimensional discrete map and perform this transformation adaptively in a topologically ordered fashion. Therefore, we set up our SOM by placing neurons at the nodes of a one or two-dimensional lattice. Higher dimensional maps are also possible to compute, but are not so common. The neurons become selectively tuned to various input patterns (stimuli) or classes of input patterns during the process of competitive learning. The locations of the tuned neurons (i.e., the winning neurons) become ordered and a meaningful coordinated system for input features is created on the network. The SOM consequently forms the required topographic map for the input patterns.

The self organization process involves four major components: (a) The Initialization. All the connection weights are initialized with small random values. (b) Competition. For each input pattern, the neurons compute their respective values of a specific function which provides the basis for competition. The particular neuron with the smallest value of the specific function is declared the winner. (c) Cooperation. The winning neuron determines the spatial location of a topological neighborhood of activated neurons, thereby providing the basis for cooperation among neighboring neurons. (d) Adaptation. The activated neurons decrease their values of the discriminant function concerning the input pattern through suitable adjustment of the associated connection weights. The flow on effects is that the response of the winning neuron to the subsequent application of a similar input pattern is enhanced. One of the most exciting aspects of SOM is that they learn to classify data without supervision. In supervised training techniques such as backpropagation where the training data consists of vector pairs (an input vector and a target vector) increased awareness is necessary. With this approach, the network (a multilayer feedforward network) presents a multidimensional input vector, and the output of this iteration is compared with the input of the next layer in the following interaction. If they differ, the weights of the network are altered slightly to reduce the error in the output. This modification is repeated many times and with many sets of vector pairs until the network gives the desired output. Training a SOM requires no target vector. The SOM learns to classify the training data without any external supervision whatsoever. For our model each quadrille our neurone defines a trajectory prototype of our input.

2.4 T-Distributed Stochastic Neighbor Embedding

T-Distributed Stochastic Neighbor Embedding is a visualizer of high-dimensional data. It is important in many different domains and deals with data of widely dimensionality.

For example, the pixel intensity vectors used to represent images or the word-count vectors used to represent documents, since typically this features have thousands of dimensions. Dimensionality reduction methods convert the high-dimensional dataset $X = \{x_1, x_2, ..., x_n\}$ into two or three-dimensional data $Y = \{y_1, y_2, ..., y_n\}$ that can be displayed in a scatter-plot. We refer to the low-dimensional data representation γ as a map, and the low-dimensional representations y_i of individual data points as map points.

Dimensionality reduction aims to preserve as much of the important structure of the high-dimensional data as possible in the low-dimensional map. The differences between various dimensionality redundancy techniques focus on what they preserve. Traditional dimensionality reduction techniques such as Principal Components Analysis (Hotelling, 1933) and classical multidimensional scaling (Torgerson, 1952) are linear techniques that focus on keeping the low-dimensional representations of different data points far apart. For high-dimensional data that lies on or near a low-dimensional non-linear manifold, it is usually more important to keep the low-dimensional representations of very similar data points close together, this is typically not possible with linear mapping.

Stochastic Neighbor Embedding (SNE) starts by converting the high-dimensional Euclidean distances between data points into conditional probabilities that represent similarities. The similarity of datapoint x_j to data-point x_i is the conditional probability, $p_{j|i}$, that x_i would pick x_j as its neighbor if neighbors were picked in proportion to their probability density under a Gaussian centered at x_i. For nearby data-points, $p_{j|i}$ is relatively high, whereas for widely separated data-points, $p_{j|i}$ will be almost infinitesimal (for reasonable values of the variance of the Gaussian, σ_i). Mathematically, the conditional probability $p_{j|i}$ is given by

$$p_{j|i} = \frac{exp(-\|x_i - x_j\|^2/2\sigma_i^2)}{\sum_{k \neq i} exp(-\|x_i - x_j\|^2/2\sigma_i^2)} \tag{2.11}$$

Where σ_i is the variance of the Gaussian that is centered on data-point x_i. The method for determining the value of σ_i is presented later in this section. Because we are only interested in modeling pairwise similarities, we set the value of $p_{i|i}$ to zero. For the low-dimensional counterparts y_i and y_j of the high-dimensional data points x_i and x_j, it is possible to compute a similar conditional probability, which we denote by $q_{j|i}$.

2.5 Affinity Propagation

Affinity Propagation (AP) clustering (Frey & Dueck, 2007) is a fast clustering algorithm used especially in the cases of large numbers of clusters. AP works based on similarities between pairs of data points (or $n \times n$ similarity matrix S for n data

points) and simultaneously considers all the data points as potential cluster centers (called exemplars).

In the AP clustering algorithm, there are two important concepts: the responsibility $R(i,k)$ and availability $A(i,k)$ which represent two messages indicating how well-suited a data point is to be a potential exemplar. $R(i,k)$ is an accumulated value which reflects how well the point i is suited to be the candidate exemplar of data point i and then sends data from the latter to the former; that is, compared to other potential exemplars, point k is the best exemplar. The availability $A(i,k)$ is opposed to $R(i,k)$ and reflects how well-suited it is for point i to choose point k as its exemplars. Based on the candidate exemplar point k, the accumulated message sent to the data point i indicates that point k is more qualified as an exemplar than others.

The sum of the values of $R(i,k)$ and $A(i,k)$ is the evaluation basis for whether the corresponding data point can be a candidate exemplar or not. Once a data point is chosen to be a candidate exemplar, those other data points with nearer distance will be assigned to this cluster. The similar value between two data points x_i and x_j $(i \neq j)$ is usually assigned the negative Euclidean distance, such as $S(i,j) = -||x_i - x_j||^2$. The algorithm uses an initial value called preference, which indicates the preference that the data point can be chosen as an exemplar. It is usually set by the median(s) of all distances. The following Algorithm 1 summarizes the process:

Algorithm 1 AP.

procedure CLUSTERINGAP(S)
  $R(i,k) = 0,\ A(i,j) = 0,\ \forall i,k$
  while Until converge **do**
    $R(i,k) = S(i,k) - max(A(i,j) + S(i,j)) \mid (j \in [1,n]; j \neq k)$
    $A(i,k) = min(0, R(k,k) + \sum_j max(0, R(j,k))),\ \mid (j \in [1,n];\ j \neq i;\ j \neq k)$
    $A(k,k) = \sum_i max(0, R(i,k)),\ \mid (i \neq k)$
  end while
  return $Trks$
end procedure

The algorithm iterates until either the cluster boundaries remain unchanged over a number of iterations or after some predetermined number of iterations. The exemplars are extracted from the final matrices as those whose "responsibility + availability" for themselves is positive.

2.6 Polinomial Interpolation

In the mathematical field of numerical analysis, interpolation is a method of constructing new data points within the range of a discrete set of known data points. In this meaning, the Polynomial Interpolation is an interpolation of a given data set by the polynomial of lowest possible degree that passes through the points of the given dataset. Let us to define it as:

Definition 2.6.1. *Given a set of $n+1$ data points (x_i, y_i) where no two x_i are the same, there is a polynomial p of degree at most n with the property $p(x_i) = y_i$, $i = 0, \ldots, n$.*

The theorem states that for $n + 1$ interpolation nodes (x_i), polynomial interpolation defines a linear bijection. Suppose that the interpolation polynomial is in the form

$$p(x) = a_n x^n + a_{n-1} x^{n-1} + \cdots + a_2 x^2 + a_1 x + a_0 \tag{2.12}$$

The statement that p interpolates the data points means that

$$p(x_i) = y_i \qquad \text{for all} i \in \{0, 1, \ldots, n\} \tag{2.13}$$

If we substitute in Equation 2.12, we get a system of linear equations in the coefficients a_k. The system in matrix-vector form reads

$$\begin{bmatrix} x_0^n & x_0^{n-1} & x_0^{n-2} & \ldots & x_0 & 1 \\ x_1^n & x_1^{n-1} & x_1^{n-2} & \ldots & x_1 & 1 \\ \vdots & \vdots & \vdots & & \vdots & \vdots \\ x_n^n & x_n^{n-1} & x_n^{n-2} & \ldots & x_n & 1 \end{bmatrix} \begin{bmatrix} a_n \\ a_{n-1} \\ \vdots \\ a_0 \end{bmatrix} = \begin{bmatrix} y_0 \\ y_1 \\ \vdots \\ y_n \end{bmatrix} \tag{2.14}$$

We have to solve this system for a_k to construct the interpolant $p(x)$. The matrix on the left is commonly referred to as a Vandermonde matrix. The condition number of the Vandermonde matrix may be large, causing large errors when computing the coefficients a_i if the system of equations is solved using Gaussian elimination.

Several authors have therefore proposed algorithms which exploit the structure of the Vandermonde matrix to compute numerically stable solutions in $O(n^2)$ operations instead of the $O(n^3)$ required by Gaussian elimination. These methods rely on constructing first a Newton interpolation of the polynomial and then converting it to the monomial form.

Alternatively, we may write the polynomial immediately concerning Lagrange polynomials:

$$\begin{aligned} p(x) = \quad & \frac{(x - x_1)(x - x_2) \cdots (x - x_n)}{(x_0 - x_1)(x_0 - x_2) \cdots (x_0 - x_n)} y_0 + \\ & \frac{(x - x_0)(x - x_2) \cdots (x - x_n)}{(x_1 - x_0)(x_1 - x_2) \cdots (x_1 - x_n)} y_1 + \\ & \qquad \ldots + \\ & \frac{(x - x_0)(x - x_1) \cdots (x - x_{n-1})}{(x_n - x_0)(x_n - x_1) \cdots (x_n - x_{n-1})} y_n \end{aligned} \tag{2.15}$$

$$p(x) = \sum_{i=0}^{n} \left(\prod_{\substack{0 \le j \le n \\ j \ne i}} \frac{x - x_j}{x_i - x_j} \right) y_i \tag{2.16}$$

For matrix arguments, this formula is called Sylvester's formula and the matrix-valued Lagrange polynomials are the Frobenius covariants.

Some applications to Polynomial Interpolation can be used to approximate complicated curves, for example, the shapes of letters in typography. One relevant application is the evaluation of the natural logarithm and trigonometric functions. Creating a lookup table and interpolating between those data points can results in significantly faster computations. Polynomial interpolation also forms the basis for algorithms in numerical quadrature and ordinary numerical differential equations and Secure Multi-Party Computation.

Note that fitting polynomial coefficients is inherently badly conditioned when the degree of the polynomial is large or the interval of sample points is poorly centered. The quality of the fit in these cases should always be checked. When polynomial fits are not satisfactory, splines may be a good alternative. However, splines generate multiple coefficients and this characteristic does not work appropriately for our approach, since the number of points varies for each trajectory.

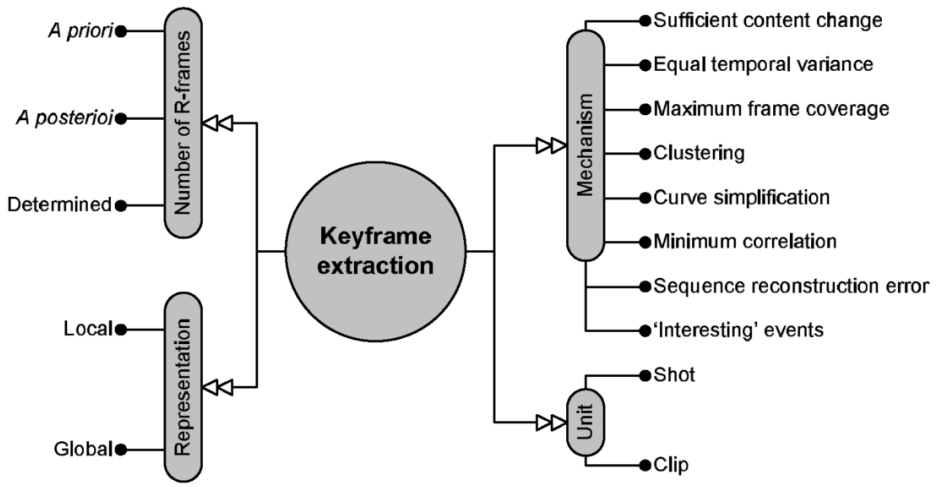
Chapter 3

Related work

In this chapter, we describe some approaches related to our proposal found in literature. One goal of video summarization is to reduce the work of the security camera operators. Nowadays, video summarization methods are applied in different fields like sports (Khan & Pawar, 2015), news, movies, series, home videos and rush videos among others. Some examples of works that perform summarization for a generic domain are Furini *et al.* (2008); Kloss *et al.* (2015); De Avila *et al.* (2011). Video summarization is an active field of research, and it is an area that has a lot to explore. According to literature, most of the works in video summarization are performed in agreement with pre-defined domains which are called structured summarization (Xu *et al.*, 2009). The other summarizers that do not depend on the structure of the content domain perform redundancy elimination. The problem with this type of summarizers is that it can not be used for any video, being that these methods are considerably limited. The domain of a video summarizer refers to what type of video is analyzing; in other words, the description of specific event, action or phenomenon. The huge number of techniques found in scholarly articles to perform summarization depend on the domain; this means that, they take into account characteristics that highlight particular properties for each type of video, for example people, cars, constant change of pixel status, sudden background changes, etc. The study of domains in video summarization demonstrate a strong relationship between specific characteristics and the content of the videos, all this to achieve better results.

According to literature on video summarization, algorithms are classified into three categories: (a) fast forwarding, (b) key frame selection and (c) frame re-composition. These concepts are briefly described in the next paragraphs.

Fast forwarding, this type of summarizer consists of skipping frames at defined or adaptive intervals; however, if these intervals are very long the summary loses semantic information. Ji *et al.* (2010) presents an approach where the local sampling rate is directly proportional to the amount of visual activity. They use clustering methods to classify the events. In this type of summarization, the video is played at a higher speed when the video content has low interest. When some frames are of interest, the video is returned to a slow rate. In the summarization of videos, the recognition ac-



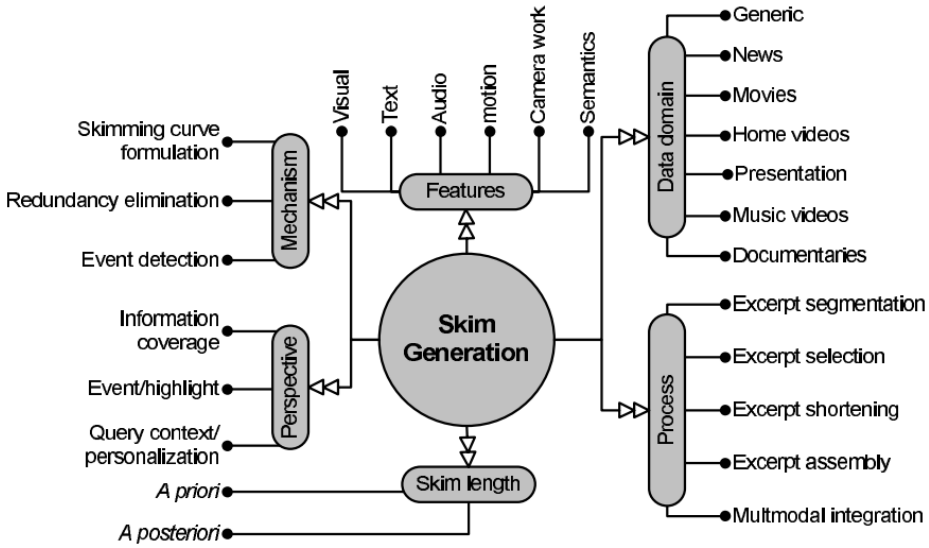
(a)

Figure 3.1: Attributes of key-frame extraction technique. The main aspects of the current approaches in key-frame extraction summarization is defined in Truong & Venkatesh (2007). These aspects are, the size of the key-frame set, the base unit, representation scope, the underlying computational mechanisms, and visualization method for extracted key-frames, the image extracted from Truong & Venkatesh (2007).

tivity (Zhu *et al.*, 2013) makes a contextual analysis of the scene. According to it, one activity does not occur in an isolated form, and it can serve as the context for others, the disadvantage is that it requires the prior knowledge of all the available activities in the dataset. According to this classification, our model belongs to this type of summarization. The difference is that we present segments of videos defined as important without variation of speeds.

Key frame selection, this type of summarizer is widely used. The simplest method for key-frame extraction is to extract examples of frames uniformly. Although this is very efficient and simple, this technique often generates redundancy in choosing key-frames and this error produces a lower representation of the video. Figure 3.1 shows the main aspects of this type of summarization defined by Truong & Venkatesh (2007). Key-frame selection retrieves the most important frames in a sequence of one video; other approaches based on these frames create dynamic summarization that consists of building shots around the key-frames for concatenating them together. In Kloss *et al.* (2015); De Avila *et al.* (2011) they perform the summarization process based on histograms of color. In particular Kloss *et al.* (2015) use Partial Least Squares Image Clustering (PLSIC), this work is getting a better result than using K-means. Part of their process consist of extracting one matrix of two hundred and fifty-six dimension to convert it to N-dimensional. This work study improves the representation of features with PLSIC reducing the dimensionality of the generated matrix. This optimization requires shorter processing time and consequently solves the problem of repeated frames. Some disadvantages with key-frame selection are the modification of the frame sequences and as a consequence the loss of the contextual information.

Frame re-composition, in this classification summarizers rebuild a new spa-



(a)

Figure 3.2: The image shows the attributes of video skimming techniques. Six aspects are taken into account to make dynamic summarization. The first is the duration of a summarization. The data domains that refer to the content of each video with this information, are: sports videos, news, movies and surveillance among others. The features that the summarizers use are audio, text, movement or semantic information. The process that the automatic summarization of video follows. Two additional aspects are perspective and mechanics. Image extracted from Truong & Venkatesh (2007).

tiotemporal segment and combine them into just one spatiotemporal scene. One characteristic of this technique is that individual frames are unpreserved as key-frame selection or as fast forwarding approaches. Some disadvantages with Frame re-composition is that the simultaneous display of multiple activities may produce confusing summaries (Lai *et al.*, 2016).

In general terms, two decisions must be made before making a summary. The determination of the summaries duration and the domain for input videos. These two decisions have a direct influence on the perspective and the type of mechanism used in the building of the summary. This influence is explained in more detail in the paragraphs corresponding to perspective and mechanisms. The determination of time in videos summaries have two types, (i) the *priori* that consists of defining intervals of time manually, it is defined before the process of summarizing and (ii) the *posteriori* that defines the intervals according to the content of the video. In these types of methods, the semantic content, the sound characteristic and also our model comes to work with them. It means that our model defines the duration of the summary once the process of video recovery finishes and our approach can reduce the amount of time to the type of information that is required to extract, which in this case is the rarity.

The video summarization is classified as video abstraction or video synopsis by other research works. The classification in three categories corresponds to the recent literature, finding few related study works under this paradigm. Traditional studies

found on literature, classify the summarization of videos into two main types, the video summarization denominated key-frames and video skimming. To contextualize our work we take these two classifications that have numerous related works. Due to our model we focuse on the summarization of video skimming. In our literature review, we also focus mainly on this type of summarization.

The key-frames is also known as static summarization which consists of the extraction of mainframes that represent the whole content of the video. Figure 3.1 shows the concepts involved in this summarization. The other denomination of this category is *Key frame selection* briefly explained in the previous paragraphs.

Video skimming is also called dynamic summarization which consists of recovering segments of videos that represent the whole content of the video to be summarized. Unlike static summarization the video segments have movement. One favorite kind of video skimming in practice are movie trailers. The Figure 3.2 shows an illustration of the techniques of dynamic summarization used. Throughout literature we see six main aspects in summaries of videos according to Truong & Venkatesh (2007): The content of the video to be summarized, the determination of the duration of a summarized video, the steps followed to perform the video summarization, the mechanism of summarization, the characteristics used and the perspective of generation. In general terms, the processes for generating dynamic summary videos can be generalized in five steps: The segmentation, the selection, the delimitation, multimodal integration, and the presentation. Some of these can be combined or suppressed. A brief description of each step is presented in the next paragraphs.

The segmentation consists of dividing the video into its fundamental expression (e.g. frames). This process is essential in dynamic summarization. It is applied to define the video units, and some segments are selected to compose the final summary. In the segmentation step, one of the most common methods is to divide the videos into segments of arbitrary size. Another older method of segmentation takes into account the variation of movement. Since the methods of summarization depends on the type of video, the content of the video is taken and highlighted features are used to distinguish between different events according to the proposed goal. In many cases, segmentation helps in obtaining good results. For instance, in the SSIG-database, which is a database originally created for anomaly detection that will be used for our experiments, the segmentation is done by movement; Chapter 5 explain this in more detail. Segmentation helps to avoid redundant video segments in which the images are static and do not contain people.

The selection consists of the application of a method or heuristic that permits the automatic process of selecting segments of videos obtained previously. After *the segmentation* is *the selection*. This is one of the most critical parts of a summarization, because the content of the clips or video segments defines the quality of the summary. For instance, in static summarization there are different methods for selecting frames. The selection of video segments defines the content and quality of summary. For example, the study work of Ajmal *et al.* (2017) uses a *curve simplification* algorithm to create an approximation of curve with less number of vertices, for then uses these

curve points as selected frames for their static summarization. Other approaches use formulas to choose segments where it takes into account the presence of audio objects or additional information. Other techniques use object detection, amount of motion, supervised classification and bio-inspired algorithms among others.

The delimitation is the next process after the video segments are selected. Usually, the critical segments are delimited. This is vital because if a mistake is made, it could lead to the loss of information producing inappropriate cut-off intervals in a summarization. To delimit the video segment the easiest way is to select a constant percentage of the segment video. Other methods include a process using accelerated adaptive suppression which consists of deleting clips that are not important and displaying the frames at a constant speed.

The multi-modal integration consists of re-aligning the delimitations of the segments that form the video summary. This process is essential because when done correctly, it can improve the coverage, the context and the consistency of the dynamic summary. For example, the study work of Evangelopoulos *et al.* (2013) integrates signals of video and audio with semantic cues (linguistic/textual) on hierarchically mode. Analyzing each modality independently for then combine features to obtain the summary. The summarizers focus on this step are based on an integration of audio, video or other similar features. They belong to the classification of the synchronized type. It means, the majority of these study works have a chronological video sequence. This process synchronizes the audio and video at a particular time. Not synchronized summaries have different methods to fuse the audio and video. The most simple integration method used in scholarly articles put the selected segments that make up the summary in chronological order.

The presentation is the final process of video summarization and consist of showing the final result obtained by the model on one screen. Literature does not focus on this process because it is manageable.

Following are the steps for video summarization. Scholarly articles present a variety of features utilized to make summaries. Their selection process is important and has a direct relationship with the quality of the final result. These two concepts also depend on the video domain type. The first step is the extraction of image characteristics with a descriptor. There are many image descriptors depending on the approach presented in literature. For instance: Scale-Invariant Feature Transform (SIFT) (Lowe, 2004) and Histogram of Oriented Gradients (HOG) (Dalal & Triggs, 2005) are two-dimensional descriptors. Histograms of Optical Flow Orientation and Magnitude (HOFM) and Histogram of Oriented Optical Flow (HOOF) (Chaudhry *et al.*, 2009) utilize temporal characteristics. In scholarly articles, these features are extracted by a descriptor and are manipulated to form a vector of characteristics. For instance, in our model, each component of the feature vector depends directly on the trajectory descriptor. Other studies in literature choose a subset of features as part of their research. These significant features are chosen to appropriately highlight characteristics of the video data which varies depending on the goal to be achieved (Li & Allinson, 2008). The features can be represented as points in multidimensional space; normally they have

changes of value according to the scenes. In our model, this type of image descriptor is used in the detection of people. The descriptor that is part of our objectives is a descriptor of trajectories created by people on surveillance videos. Our model performs a complete pre-processing to improve the trajectories clustering. These visual features are widely used in literature and generally employed to measure the similarity between frames. These characteristics are commonly used by summarizers to eliminate redundancy. Cahuina (2013); De Avila *et al.* (2011) present local color histograms to perform the detection and the composition of video segments. According to Khan & Pawar (2015), in sports videos, the visual elements such as color, edges, texture and their spatial position play an essential role in the identification of important events which are then utilized to estimate essential events. Text is another feature that has been used in video summarization. Text can be automatically extracted from a video with a text extractor or manually via an audio transcription or other external sources. Text features are substantially related to the audio of the video. The text features can improve the semantic extraction on one video when it is compared to feature extraction using only the video and the audio. The text can be obtained directly from the video stream as legends. Typically, text information is used when searching keywords that are important indicators in an event. The audio of a video can be described as audio features. These are associated with a specific type of video and can detect interesting events, for instance, the word "goal" in soccer videos. The visual dynamic characteristics are based on the activity in the scene. In other words, the larger amounts of activity presented in a scene will result in more data. This information is commonly used for behavior detection. Camera movements are generally used in sports events and rush videos since the camera movement is strongly related to these types of videos. Finally, the semantic characteristics refer to the meaning of a scene in videos. Semantic characteristics in the process of summarization are strong indicators of different events (Cunha, 2011).

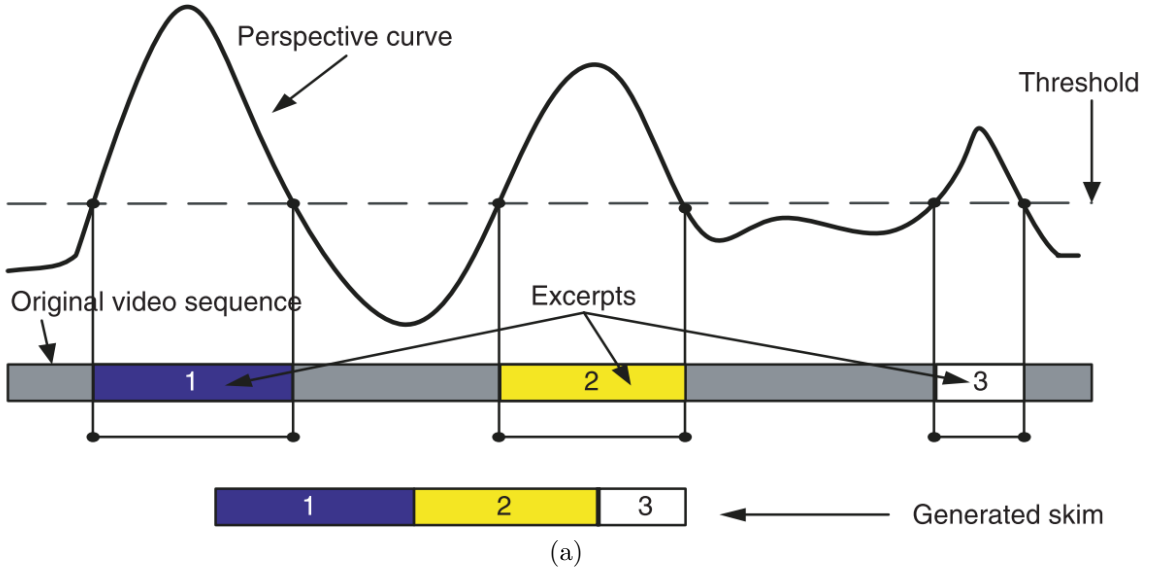
The *preserved perspective* must be according to the objective of the summary since different perspectives create different summaries. According to Truong & Venkatesh (2007) preserved perspective is classified in three ways: (a) *Coverage of information*, helps in the elimination of redundancy. it also focuses on the preservation of information and content coverage. In other words, the complete content of the video is shown without affecting the understanding of the observer. In this approach the observer wants to have the whole meaning of the video. (b) *Interesting or important events*. This technique summarizes by importance. The concept of important varies depending on the type of video which the observer is interested in. It is applied to videos where the concept of interesting can be defined, modeled and extracted. In other words, this applies to videos that have structured content and one defined domain. One of the main problems with this perspective is to define the limits of the detected events, guaranteeing consistency and preserving the context of the video. (c) *Query context and customization*. This perspective is based on inquiries or personalization of context. It examines the request for information of the observer for which a summary is made. In other words, it is based on queries from the observers who provide a preference that is taken into account to make a summary. For example, the word "goal", the sound of a gun or the pronunciation of the name of some favorite soccer player. Another method

of this approach is the adaptation of personality, or according to the preference of the observer. This summary is different according to the preferences of the observer. For example, if it is a man, a woman, a child, an adult, a researcher or an athlete the video summary should reflect their objectives.

The *mechanism* depends and is defined from the preserved perspective. For example, the summaries that aim at the detection of specific events will lose the aspect of coverage a video, and consequently, the method used would suffer a change in its mechanism. The mechanism can be ordered in three types. (i) *Elimination of redundancy* consists of deleting segments that contain similar content. These divisions can be done in the segmentation or delimitation process of the summary. For instance, in surveillance video one approach to redundancy elimination is made by motion detection. It deletes any segment that does not contain movement. Another example is the detection of segments that do not contain noise and are therefore eliminated. However, this requires a video with an audio channel. This technique eliminates redundancy with the help of a number of channels. The detection will not function using a single channel (Cunha, 2011). (ii) *Detecting interesting events*, the objective is to identify and locate specific spatial-temporal patterns (e.g., a person with a card in their hand). This approach involves the detection and delimitation of some specific events. To identify these incidents it is necessary to use audio, visual or cinematographic characteristics, which are generated from important events in the video. (iii) *Summarization curve formulation*, this process is associated with a curve of perspective used in the summarization. It consists of choosing values which are above some defined adaptive threshold so that features related to scores can be created. With this type of mechanics, punctuation is based on specific characteristics referring to the perspective found. Figure 3.3 shows simple process of generating summaries based on the perspective curve. For instance, in summaries based on preferences, these descriptors have to reflect the requested preference. For example, one of the main tasks of defining the descriptors precisely in soccer is taking into account the sounds that the audience makes or the screams of "goal". Within this category, there are other descriptors for generic video. The problem with curve formulations is that it does not ensure consistency and a balanced content coverage when the video has short segments or similarity. However, there are methods that address this problem (Truong & Venkatesh, 2007).

Our approach uses unsupervised clustering to classify and detect the morphology of trajectories. In problems related to videos, summarization clustering is widely used. Some examples of this are Cahuina (2013); Cunha (2011); De Avila *et al.* (2011); Kloss *et al.* (2015); Höferlin *et al.* (2013). In general terms, they process the features extracted from images that form the videos before applying them to a cluster. Grouping features means joining similar data and separating the dissimilar data. Clustering is an active topic of research interest. Clustering is not only used with image processing because it also has applications in fields like biology, medicine, business, marketing, the world wide web, computer science, management, statistics, pattern recognition and social science among others.

There are different clustering paradigms in literature and each of these classify clustering in different ways. For instance, according to Tian (2015), the clustering al-



(a)

Figure 3.3: Summary from a curve. Shows a straightforward process for summary generation based on the perspective curve, the image extracted from Truong & Venkatesh (2007) study.

gorithm could be viewed from two perspectives, the traditional and the modern: The traditional clustering algorithms have nine categories with twenty-six algorithms while the modern clustering algorithms can be divided into ten categories which mainly contain forty-five commonly used algorithms. Other paradigms found in literature arrange clustering in five categories, (i) the clustering based on connectivity also called hierarchical clustering, (ii) the clustering based on centroid, in which the algorithms used are Kmeans (Jain, 2010), Kmeans++ (Arthur & Vassilvitskii, 2007) and the affinity propagation (Frey & Dueck, 2007). (iii) Clustering by distributions like EM-Clustering algorithm (Do & Batzoglou, 2008), (iv) clustering based on density (Chen *et al.*, 2017) and (v) specialized types of group analysis like the spectral grouping algorithm using spectral graphs (Vázquez-Martín & Bandera, 2013). Clustering methods have advantages and disadvantages depending on the scenario in which these are implemented. In our study clustering plays an important role since it can solve the problem of detecting uncommon features in trajectories. Using clustering methods on surveillance videos that include trajectories is a field yet to explore. Clustering can provide evidence that can classify different patterns formed by each event that happens in surveillance domain (Damnjanovic *et al.*, 2008; Höferlin *et al.*, 2013; Ji *et al.*, 2010; Pritch *et al.*, 2009; Wang *et al.*, 2011).

Finally some abnormal trajectories detection and trajectory descriptors works are discussed, as well as the databases that they use to experiment and compare their results. The description or modeling of trajectories is the first step in the treatment of trajectories. Kong *et al.* (2018) presents a classification of trajectories as well as a summary of applications and services that use trajectories. It defines Explicit trajectory data as trajectories that are related to time and place, the Global Positioning System (GPS) generated trajectories are the most popular in this classification. This study work also defines Implicit Trajectory Data like all other trajectories that are not defined

as Explicit trajectories, the survey presents a classification based on the applications that give trajectories and mention some services of recommendation systems which use trajectories in their studies.

Since our approach are related with rare trajectories detection, we mention some studies related on trajectory anomaly detection. Piciarelli *et al.* (2008) presents results on anomaly trajectory detection using Support Vector Machine (SVM). For this study work Piciarelli create an algorithm which generated a Synthetic dataset, which consist on trajectories of 16 points generated automatically. They made multiple experiments highlighting those with better results. Laxhammar & Falkman (2014) presents improvement results of Piciarelli study work. They work on design a sequential analysis of incomplete trajectories or online learning based on an incrementally updated training set. They implement and propose the Sequential Hausdorff Nearest-Neighbor Conformal Anomaly Detector (SHNN-CAD) for online learning and sequential anomaly detection in trajectories achieving competitive classification performance with minimum parameter tuning. Ergezer & Leblebicioğlu (2016) presents a trajectory descriptor with covariance matrix for detect anomaly trajectories use nearest neighbors and for Activity Perception use spectral clustering. This study works using the Synthetic dataset created by the algorithm of Piciarelli and for real dataset use UCSD anomaly detection dataset and MIT Parking Lot dataset. Sillito & Fisher (2008) proposes a new framework to detect abnormal trajectories. This considering the behavior of passersby in terms of trajectory motion. The framework build a one-class training that is based on probabilities using the Gauss distribution. This study use to represent a trajectory the B-spline curves, for this purpose is necessary to transform a trajectory in two subtrajectories (coordinate x or y vs time coordinate). The framework makes a learning process using both tagged and untagged data and uses two databases for experiments. The first is CAVIAR "INRIA" Dataset and the second is Carpark Dataset.

Some final considerations of this Chapter are: The mention of concepts involved with the video summarization as an introduction to the area, the mention of how video summarization is present in the literature, the review articles published in video summarization in different domains, academic works that process surveillance videos, clustering methods and study works related to the processing of trajectories. The conclusion of this Chapter is that the literature not present deep studies in the field of video summarization with trajectories. These work studies serve as a reference since we use the detection of rare trajectories in our approach.

Chapter 4

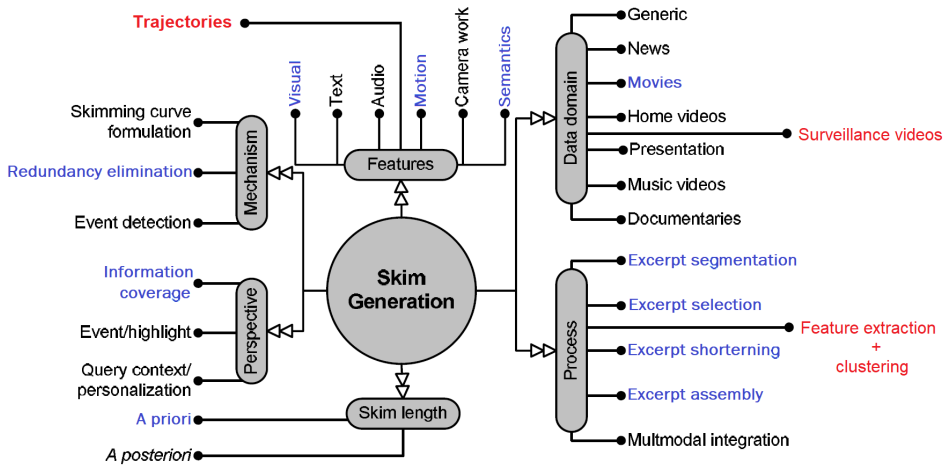
Methodology

4.1 Overview

This chapter explains our approach for video segmentation based on the trajectory information. The scope of our approach are surveillance video from fixed cameras. The proposed method focuses on the extraction of semantic information from trajectories, to differentiate current trajectories and also to point out the rare trajectories. Surveillance videos have a particular context; for example, they contain recurrent scenes where determinate events happen constantly. In many cases, these surveillance videos contain long periods of uninteresting events, there are some types of videos in which people are constantly present in a scene, is for this reason that is possible to considerate the presence of people as a factor of interest or importance for the spectator. Our summaries rely on the context; to put it differently, it highlights the behaviors that happens in less quantity and repeated paths permits to separate rare events. This characteristics are considered for gathering a summary.

Figure 4.1 depicts an overview of the proposed methodology and also draws the relationship between our model with methods presented in literature. The contributions made in the present study are colored in red, and the traditional approaches used in the literature are in blue. Our approach shows the functionality of our straightforward model to summarize videos in surveillance scope. We evaluated our model with an original dataset from laboratory surveillance camera view.

Figure 4.2, presents the pipeline of our approach. Our model begins detecting people in a video sequence, after that it builds the tracks for each of them. The trajectory generation looks for a specific body point of people. To define reference point we employ a person pose estimation (Cao *et al.*, 2017). Afterward, reference point is used to create the tracklets using a heuristic based on association data estimation. At this moment it is essential to remember that each trajectory corresponds to one detected person. Our model describes the people trajectory using their morphology to create the feature vectors. Next, our model defines a temporal sub-series extracted



(a)

Figure 4.1: Attributes of video skimming techniques and contextualization with our approach

from spatial location of trajectories. These two sub-series are combined with other denominated time components. One by one, each unit of time is processed until the total number of points of this trajectory are identified. This decomposition is computed to improve the trajectory modeling. With this information, we can apply a different method to describe each sub-trajectory and as a result obtain sets of characteristics. Once the model has these characteristics, the clustering can be performed directly on them, and with the output of clustering, the model obtains subgroups of trajectories. In these subgroups, it is possible to differentiate paths by their shape and behavior. Then, depending on the number of elements they have, it is possible to label them as rare or normal. Hence, we can define the rarity in each cluster as our hypothesis mentions: "rarity will be inversely proportional to the number of elements of each cluster ". In this study, we want to prove that trajectories provide knowledge about the standard or rare motion of people. At the same time, it can answer queries about how rare a video is based on the trajectories that cluster contains. Finally, videos can be summarized according to their highlights of unusual events based on paths.

4.2 Approach

In this section, we expose our pipeline as well as a detailed description of our proposal. The pipeline is divided into three parts: *preliminaries*, *feature extraction* and *trajectory analysis*. In the following subsections we describe each of these parts with more detail:

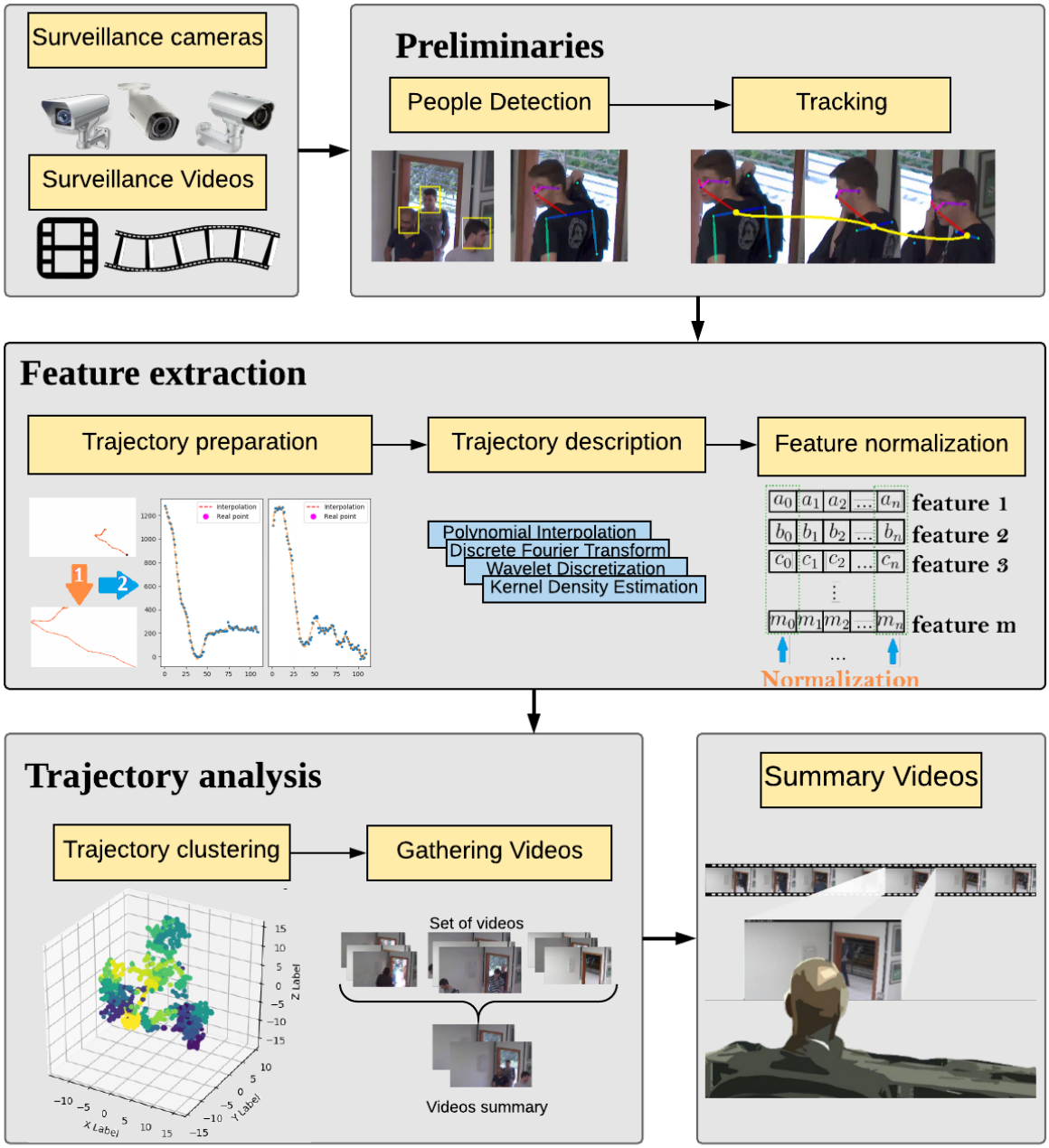


Figure 4.2: Illustrative diagram of overall pipeline.

4.2.1 Preliminaries

In *preliminaries* we describe two techniques used for the extraction of trajectories. The input of our model is composed of surveillance videos, and trajectories are generated for each person. This process is divided into *Pose estimation* and tracking. The *Pose estimation* compute a silhouette of a detected person (Cao *et al.*, 2017) while tracking is used to find the correspondence of new detected points. Our tracking method uses the Kalman filter (Weng *et al.*, 2006). The combination of these two preliminary steps allows us to obtain defined trajectories. The two mentioned methods are described briefly in the following paragraphs.

Multi-Person Pose Estimation. The two dimension pose estimation solves the problem of localizing anatomical key-points or parts of the body for individuals. The information that we get from the detection of people is not enough to have a fixed point for the monitoring made for each person. One person can create different shapes of trajectories depending on the reference point (Liu *et al.*, 2009). For our study, we consider as a trajectory point the *reference point*. Figure 4.3b shows an example of this. The reason for using Pose Estimation is that many detection algorithms and also tracking algorithms use enclosing boxes.

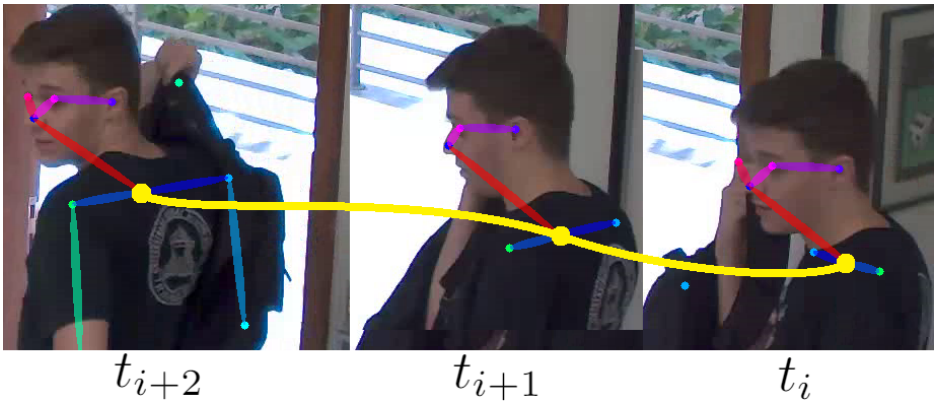
This is used to enclose a region in which a person is located and has a variation in size from frame to frame. For instance, if a person stretches their hand, then the enclosing box increases its dimension causing a loss to precision of the reference point. Using Pose Estimation increases the accuracy of reference points. An example of pose-estimation is shown in Figure 4.3a.

(a) (b)

Figure 4.3: The figures show the results of Pose Estimation, in (a) the silhouette obtained once applying pose-estimation and in (b) the point of the body chosen to generate our trajectories (fiducial or reference point).

People tracking is done in chronologically arranged frames. It can be difficult to perform when the speed of the target is high and when the target changes its direction. A tracker aims to generate the trajectory of an object over time by locating its position in every frame. Some types of tracking are *point tracking*, *kernel tracking* and *silhouette tracking* (Yilmaz *et al.*, 2006). We use Kalman filter tracking algorithm (Welch *et al.*,

1995), that fall in the category of point tracking. Kalman filter is a deterministic and statistical model, which estimate the state of a system, where state is Gaussian in its nature. The tracking algorithm use a heuristic approach that links tracklets in previous frames with the current frame. Based on data association, it attempts to build the tracklet for all the individuals in the scene using the movement information. Thus, each point is related to some new point depending on their velocity and orientation. In consequence, for each frame, this algorithm employs the Kalman filter estimation and the new point creates a score. The scores for all points are them stored in a matrix. Finally, the point associated with the corresponding tracklet is entered using the Hungarian matrix algorithm. The tracking model proposes to use the area of one found object in the following frames. To link it with a certain amount of pixels, will depend on the movement of the object. This process is performed in two consecutive frames. The videos can present total or partial occlusion of the object. Tracking then defines its different levels (Weng *et al.*, 2006). With tracking it is possible to describe the movement translation of a person. Figure 4.4 shows an example of this.



(a)

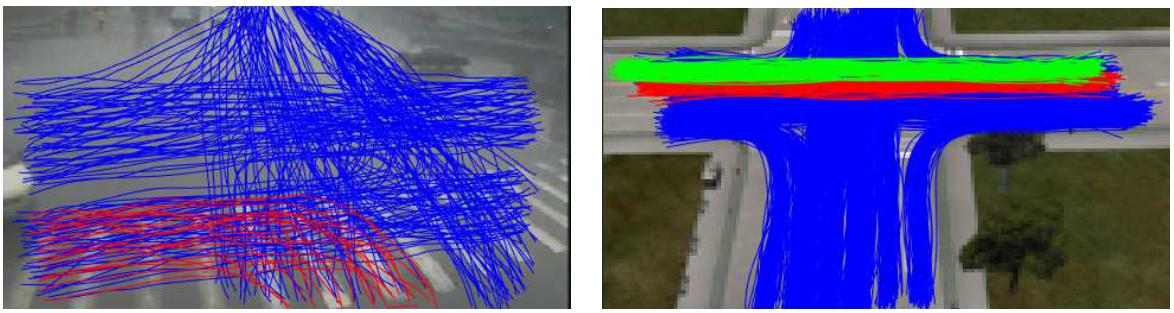
Figure 4.4: For our study, we take the midpoint of the segment that joins the shoulders of a person with the segment that represents the neck for generating trajectories (yellow line).

4.2.2 Feature extraction

This subsection is divided into three parts, which are: *trajectory pre-processing*, *trajectory description* and *feature normalization*. Using obtained trajectories in *preliminaries*, our model normalizes them to improve the results of the description, after that trajectory decomposition is applied. The next step is to compute the descriptor. For this, we use classic descriptors and also introduce a novel descriptor based on polynomial decomposition. The output of this process is a feature vector for each trajectory.

4.2.2.1 Trajectory preparation

In this subsection we describe the techniques that we use to prepare trajectories before clustering: *trajectory normalization* and *trajectory decomposition*. These previous processes are used to improve the description of the trajectories (Sillito & Fisher, 2008; Hu *et al.*, 2013; Zhang *et al.*, 2009; Naftel & Khalid, 2006). Xu *et al.* (2015) justifies the use of *trajectory preparation* clustering the trajectory points directly and obtaining as a result the Figures 4.5a and 4.5b. In this example, the experiments were performed using trajectories extracted from vehicles. They used three databases based on the fixed camera view. In both images performed in this experiment, we observed trajectories that are overlapping. These represent different routes generated by the cars and in these images we can see the obtained results after applying this technique. Although the trajectories provides the information of target location, processing these points directly provides inaccurate results. Two reasons lead to the failure of this strategy (Xu *et al.*, 2015): (a) *variations within a cluster*. Due to the uncertain movement of people, it generates a variety of trajectory forms. This leads to inaccuracy in the extraction of trajectories. In Figure 4.5a the trajectories colored in red belong to the same cluster. Nevertheless, they have different morphology. (b) *ambiguities across clusters*, are scenarios in which the similar trajectories in shape and close positions belong to different clusters due to their semantic differences (e. g. on Fig. 4.5b red and green trajectories are in different clusters, for instance, when they are located in different traffic lines). As a result, the position information is ambiguous among clusters.



(a) Variations within a cluster (b) Ambiguities across clusters

Figure 4.5: The illustration of the challenges of trajectory clustering, both images shows deficiencies when applied to the cluster directly on trajectory points. On (a) red trajectories belong to the same cluster since on (b) green and red trajectories belong to different clusters. Images extracted from (Xu *et al.*, 2015).

Trajectory normalization. This normalization aims to capture the morphology of a trajectory without taking into account the place in the image frame. Thus, the location information does not influence the descriptor. Our model focuses on the morphology and discarding the location of the trajectory. Before explaining the normalization, we will define some important concepts: *point* and *trajectory*. The next definition comes from the inspiration of a scholarly article.

Definition 4.2.1. *A point p is a tuple (x, y, t), where x and y are the position in the*

image and t is the time lapse of when the position is collected, where $k \in \mathbb{N}$.

$$p_k = (x_k, y_k, t_k). \tag{4.1}$$

A list of points ordered in time forms a trajectory.

Definition 4.2.2. *A trajectory* T_i *is a tuple (* tid_i *,* $\{p_1, p_2, ..., p_K\}$ *), where* tid_i *is the identifier and* $t_1 < t_2 < t_3 < ... < t_K$ *in a sequence of points* $\{p_1, p_2, ..., p_K\}$*, where* $\{i, K\} \in \mathbb{N}$.

$$T_i = (tid_i, \{p_k\}_{k=1:K}). \tag{4.2}$$

For our normalization technique, we use the *Feature scaling* method. It is used to standardize the range of independent variables in features of data. Since the range values of each component can vary widely, the normalization re-scales each component value between zero and one.

Let the following spaces be:

$$w_x = \{x_i \in p_i \mid \forall p_i \in T_j\}, \tag{4.3}$$

$$w_y = \{y_i \in p_i \mid \forall p_i \in T_j\}, \tag{4.4}$$

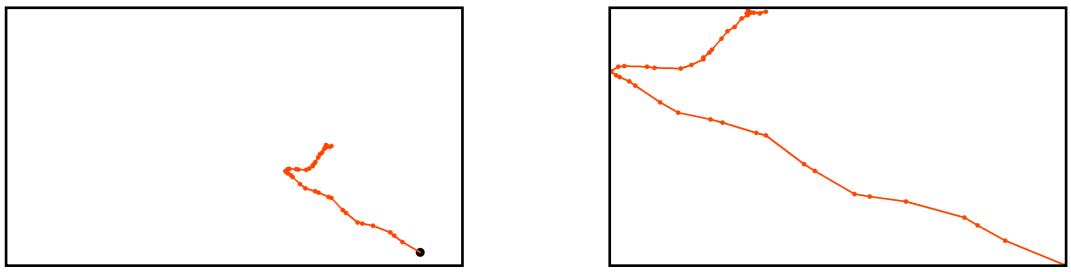
where p_i is a point and T_j is a trajectory. For all variables x_i and y_i of T_j, the following formulas are applied:

$$x'_i = \frac{x_i - min(w_x)}{max(w_x) - min(w_x)}, \tag{4.5}$$

$$y'_i = \frac{y_i - min(w_y)}{max(w_y) - min(w_y)}, \tag{4.6}$$

where *min* returns the minimum value of one space and *max* returns the maximum value of one space. After computing each component with the Equations 4.5 and 4.6, each element has a new assigned value. The value zero is the minimum. The value one is the maximum and the rest of the intermediate values are scales between those thresholds. For visualization purposes, we proceed to multiply these values by the dimensions of the original image and as a results, we obtain the visualization of the trajectory in Figure 4.6.

Trajectory decomposition. Trajectory decomposition aims to transform a trajectory into signals. The signals focus on the shape variations. This technique enables the model to describe paths omitting redundant information of each component. Before explaining the decomposition process, we will define a sub-trajectory:



(a) (b)

Figure 4.6: The normalization of trajectories improves our features. In the left figure (a) the initial trajectory corresponds to the path generated in a video segment. In right figure (b) each component of the trajectory has been multiplied by the dimensions of the video.

Definition 4.2.3. *A sub-trajectory T'_s is a set of points $p_k = (c_k, t_k)$, where t_k is the time instant in which the component c_k is collected and $c_k \in w_x$ or $c_k \in w_y$. Where w_x and w_y were defined on Equations 4.3 and 4.4 respectively.*

$$T'_s = \{p_k\}_{k=1:K}. \tag{4.7}$$

For the purpose of this study, our descriptor break the trajectories down into two set of points. We define these sets as sub-trajectories, there are two types of sub-trajectories, ones that represent the Horizontal movement (Figure a) and others that represents the vertical movement (Figure b) of the fiducial point. This decomposition allows the descriptor work on a trajectory by two parts over time thereby increasing the reliability of the trajectory description behavior.

The trajectories can be modeled as mathematical relations and after each trajectory is decomposed into two sub-trajectories, the conversion into signals is done, and therefore it can be seen as mathematical functions. In other words, our model converts one mathematical relation (trajectory) into two signals. This process aims to improve the description process, due to this model is possible to separate the relationship between vertical and horizontal movement. Also, with this decomposition, the signals accomplish the objective property of a mathematical function, in which one unique element of the domain (the time) corresponds to at least a element in the range (vertical or horizontal positions in the frames). This property simplifies and improves the description. Figure 4.7 shows an example of sub-trajectories generated by our model. Another interpretation that fits into this process is that these sub-trajectories are time series that represent the variation of each spatial component of the trajectories.

4.2.2.2 Trajectory description

The descriptor inputs are the two sub-trajectories obtained in *trajectory pre-processing*. The output of this part of the method is two feature vectors, one for each sub-trajectory.

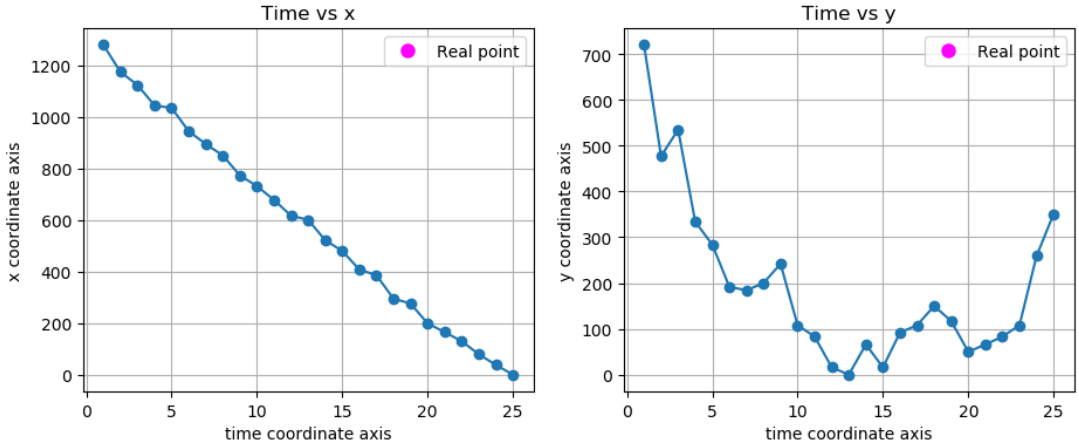


Figure 4.7: Example of trajectory decomposition.

This output is concatenating to form one feature vector. For our study the following techniques are used as descriptors:

Polynomial interpolation. The interpolation is the procurement of new points from a discrete set of points. These points are obtained from our experiments. For this purpose, a sub-trajectory could be described as a mathematical function. From N sub-trajectory points we can obtain two components that are coupled together to create c_k. The points of a trajectory can be defined as (c_k, t_k) and the function f for the interpolation is defined as:

$$f(x_k) = y_k, k = 1, ..., n \tag{4.8}$$

where sub-trajectory composition is $(x_k, f(x_k))$ each x_k are called nodes.

Polynomial interpolation is the generalization of linear interpolation where linear interpolant is a linear function. The interpolant is replaced with a polynomial of higher degree. In this part, we are modeling the trajectories points like polynomials. The coefficients describe the morphology of trajectories. This technique brings some advantages; for instance: the interpolation is independent of the number of trajectory points. This characteristic avoids the normalization of number points as a result, they have the same dimension of the feature vectors for each path. Also, the interpolation brings an interesting property; it is sensitive to variations of each trajectory position point; for instance, if there are two sets of points that differ by only one, the interpolation can describe this small change as resulting in different coefficients. For our model, we remove the independent term of the generated polynomial because this term provides information about the location in the frame image and for our model based on trajectory morphology this is not necessary. The polynomial coefficients have the following form: $P(x) = (P_n x^n, P_{n-1} x^{n-1}, ..., P_2 x^2, P_1 x)$ where each one of these coefficients can describe the morphology of the interpolation function. Figure 4.8 shows how each polynomial coefficient independently describes a different behavior. This is useful when we need to model the trajectory. The interpolation process generates a variety

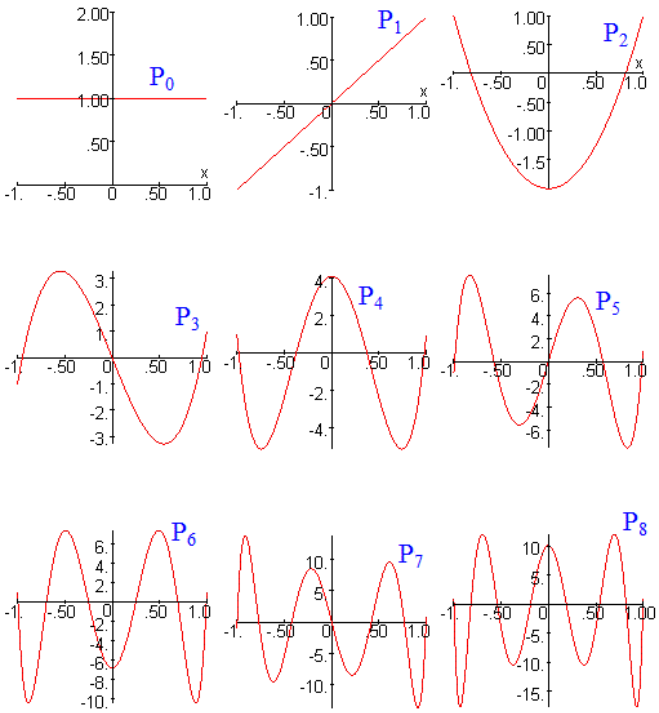


Figure 4.8: Graphic of the first nine variables of Legendre polynomial taken independently. Image extracted from(Téllez, 2009).

of polynomials and it depends on the complexity of the sub-trajectory morphology. For instance, trajectories with a more distorted morphology have higher values in its coefficients, whereas the coefficients of smooth trajectories tend to have zeros especially in the coefficients that have variables with a greater degree. It informs us that when the trajectory tends to be similar to a line, the coefficients which have variables with the low degree, will have high values.

The problem with interpolation is the approximation of one complicated trajectory by other that are more simple, Figure 4.9 shows an example of this occurrence. This Figure shows an example of applying the polynomial interpolation approach in a trajectory. Applying it in both sub-trajectories, the function obtained by the interpolation (red line) does not describe completely each real point (blue points) extracted from our experiment. The problem of inaccuracy in the polynomial interpolation was overcome with the Discrete Fourier Transform and the Wavelet Transform. Once the coefficients in each sub-trajectory are obtained, these coefficients can be used as characteristics of a vector. After concatenating these two feature vectors, the vector of characteristics is ready to be used in the clustering process.

The Discrete Fourier transform. The Fourier transform permits the transport of signals into a domain of frequency. Our descriptor apply DFT directly on the space-time sub-trajectories to obtain as an output a set composed of complex numbers. Using the absolute value of these complex numbers we can transform them into real numbers. Once having this set of values our descriptor use histograms. A histogram is an accurate representation of the distribution of numerical data. The next stage is

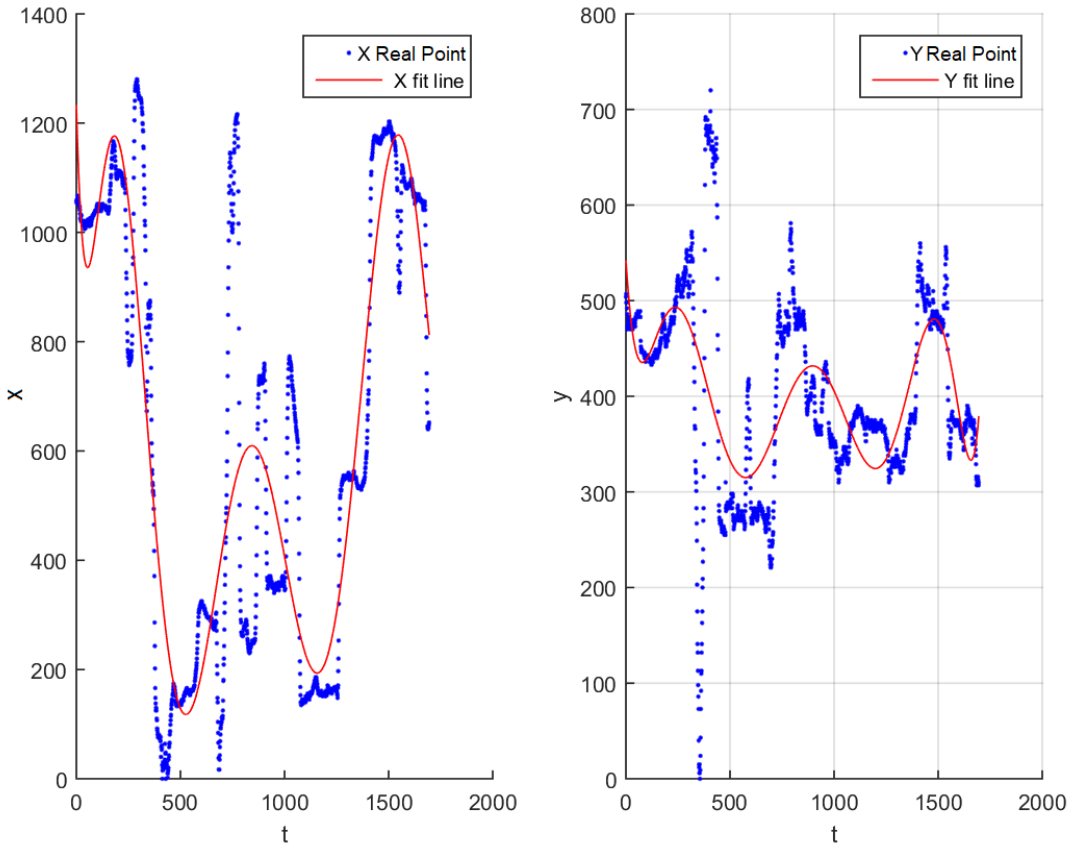


Figure 4.9: Problem fitting generated by polynomial interpolation. It is applied on one trajectory, creating two sub-trajectories. The fitting function result is plotted in red color while real points extracted from our experiments are plotted as blue points.

to introduce each value in a histogram to produce the feature vectors and constructing a representation based on distribution. We employ this technique as a descriptor because the signals of the decomposed trajectories present high variation in the range axis (movement information) especially when the trajectory has high deformation. Therefore, this type of information may be captured using the Fourier transform.

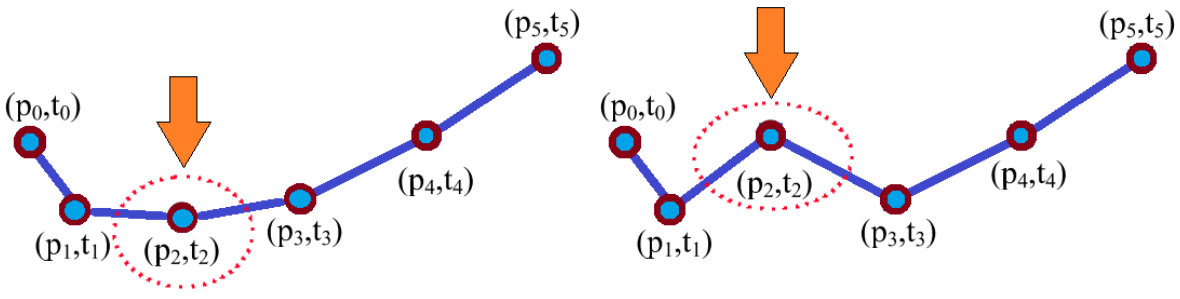
The Discrete Wavelet transform - haar. It transforms each trajectory into the frequency domain as Fourier transform does. This Wavelet transform, the haar presents different levels of frequencies depending on the number of different forms present in the trajectory. Like the Fourier transform, our descriptor constructs a representation based on distribution or histogram. The histogram is constructed with all the frequencies obtained. As a results, the model has as features the bins of the histogram. The feature vector has two sets of bins, one for each sub-trajectory. In addition to this feature vector, our model adds as a characteristic, the corresponding number of intense frequencies found by the DWT. An important characteristic of Wavelet is that these variations may appear at different times. It means, this method is appropriate because it has sensitivity to perceive different frequencies, signal variations that can be reflected

in the morphology of a trajectory. However, the trajectory may be smooth at the beginning of the event but could become disturbed in its course. Therefore the DWT detects these types of situations and incorporate this information into the descriptor. DWT was also used by Tint & Soe (2013) to summarize static videos, this study does key-frame detection using wavelet detail coefficients to represent frame content. When detail coefficients change and the subtraction overcome the threshold, the last frame of the pair is detected as key-frame by this method. Tint & Soe (2013) present a different use to the DWT, creating the antecedent as a method to built static summarization on surveillance videos.

Feature normalization. This method is used in the range of our variables of each vector. Normalization means adjusting values measured on different scales to a common scale. Normalization assigns new values to the input variables depending on the type of normalization. For our study, *max-min normalization* is used. This normalization restricts the range of values of feature space between the maximum and minimum value and assign it from zero to one. The normalization performed in each component of the feature vector, gives it a wide variety of output data that must be normalized. In our model, normalization makes comparisons between elements of each space with a good distribution.

4.2.3 Trajectory analysis

Trajectory analysis is the last part of our pipeline. The idea is to apply a descriptor using the extraction of semantic information from the trajectories. Once the characteristics of each trajectory are obtained, the model is ready for clustering. After gathering videos as output, the model provides videos chosen automatically as rare. This selection is made based on our hypothesis. In other words, Our model aims to separate uncommon videos from others, where the number of elements in one cluster can define the rarity of it. The cluster with the smallest number of rare elements is selected first and the following clusters have increased numbers of rare elements.



(a) Trajectory A (b) Trajectory B

Figure 4.10: Our approach will be sensitive to little changes in trajectory points because when one point is moved, the morphology of the trajectory changes.

Some algorithms of clustering like Traclus (Lee *et al.*, 2007) and Liu *et al.* (2014) perform a segmentation process before clustering. Our approach seeks to find and to

be sensitive to this tiny variation in trajectory morphology. One of the most relevant characteristics of our model is the modeling of each trajectory as a unique identity. By processing it completely and not by segments, this features is able to find the rarity in a set of trajectories. Figure 4.10 shows an example of the semantic information to be retrieved. We can see two trajectories that are similar; however, they are not the same. This type of information is lost when we segment the trajectories. In the example, we can see that trajectory "A" differs from trajectory "B" only at the point (p_2, t_2). When this point changes position, it modifies the morphology of the trajectory and in some cases, the rarity is observed in these changes.

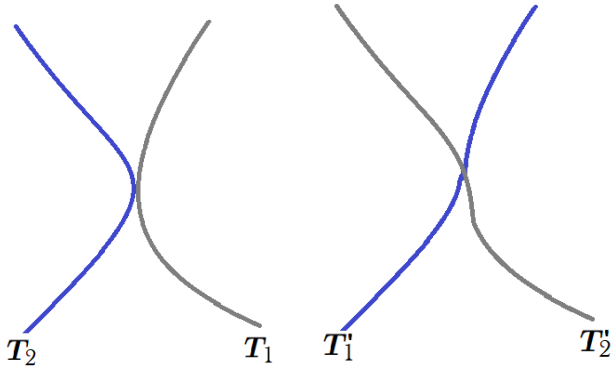


Figure 4.11: An example shows both figures (a) and (b) which could be easily confused. Figure (b) shows two trajectories that intersect while on figure (a) this does happen.

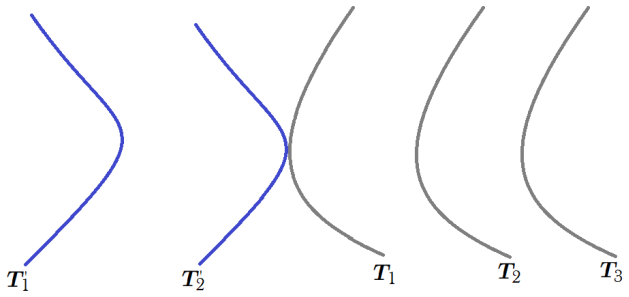


Figure 4.12: Trajectories that are placed on different location; however they have the same behavior, they can be classified as the same, it is achievable due to the process of normalization.

Figures 4.11a and 4.11b show situations in which two trajectories could be confused with each other. This confusion happens when one model processes the paths sequentially and from this orders the set of segments. Figure 4.12 shows trajectories T_1, T_2 and T_3 as the same trajectories in different positions. An algorithm that does not take into account the variations of translation could classify these trajectories as different, when in fact, it is a single pattern of paths. Our approach describes the shape of a trajectory. To achieve this, the normalization is carried out between the set of points forming a trajectory. This step is described as trajectory normalization and is explained in Section 4.2.2.2.

Some unsupervised clustering methods are used. It is important to note that unsupervised clustering is not limited to the number of clusters. Therefore, this method can generate new classes, and as a consequence, it generates new classifications. The creation of new clusters depends on the number of different shapes that our trajectories have. For example, when one new trajectory appears and it has an unusual shape, it is assigned to the new cluster. This last part of our pipeline consists of *trajectory clustering* and *gathering videos*. In the following subsections, we describe these processes with more detail.

4.2.3.1 Trajectory clustering

The automatic grouping of trajectories is fundamental in several applications such as anomaly detection (Ergezer & Leblebicioğlu, 2016; Piciarelli *et al.*, 2008; Laxhammar & Falkman, 2014), modeling, retrieval (Hu *et al.*, 2013), or in sports (Turchini *et al.*, 2015), in which there are a variety of trajectories and movement patterns. This type of clustering divides groups appropriately without the need of prior labels. Trajectory clustering methods are divided into three categories: The unsupervised, supervised and semi-supervised (Bian *et al.*, 2018) clustering methods. A crucial challenge in our approach for clustering is to define appropriate features that can separate trajectory information in the clustering process. For our clustering process we use the following methods: The Self Organized Maps (Schreck *et al.*, 2009) and Affinity Propagation. We mention these specific models because these two suit our experiments. For our study, we use SOM to group similar features by taking advantage of the function to sort properties. In our model, one neuron is assigned to one cluster. SOM as a clustering method is supervised since we define the size of the network as well as the number of nodes. Another property that SOM has is the similarity between neurons that are adjacent. This characteristic is caused due to the neighbourhood function that preserves the topological properties of the input space. In our model, each neuron represents a cluster. The net is trained with all feature samples. In the other hand, after experiment with Affinity Propagation a better result was achieved. We can show it on Table 5.1. We describe extensively the mentioned algorithm in the next paragraph since it forms part of our approach:

Affinity propagation. This clustering method uses the concept of passing messages (Frey & Dueck, 2007). Affinity propagation does not need to know the number of clusters for the process of grouping. It is a characteristic of a non-supervised clustering method. A part of this technique seeks a representative from each group. It means that the algorithm finds one representative feature vector for each cluster. For our model, one point is represented by a vector of characteristics. Descriptors that extract the features are explained in the previous section. These features are applied to the affinity propagation clustering method. Affinity propagation allows the separation of trajectories that are different without problems. The disadvantage with this technique is the creation of many clusters generating extra classes. This disadvantage depends on intraclass variability of the used dataset. For trajectory clustering, this property is inconvenient because it segments groups with low intraclass variability. However, for

our summarization method, which consists of showing with priority the unusual events (video summarization), it works adequately.

4.2.4 Gathering Videos

Finally, with clusters, we can assign more priority to videos that have rare trajectories in it or determine segments of a video where this type of paths appears. The unusual trajectory is defined as a trajectory that is in a rare cluster with few elements. In our experiments we demonstrate that the cluster with the highest number of components tends to describe an everyday event. On the other hand, groups with few elements are labeled as a rare cluster. As explained, our summarization model is based on this idea. The clusters with the number of factors less or equal to determinate threshold t, are small clusters. In our model, the threshold is essential only if it is necessary to take quantitative measurements to find the percentage of reduction of a video summary, as it is done in literature. For our approach, it is more important to show videos with precedence. The order of videos presented is more important than the percentage reduction of the video.

Chapter 5

Experiments

In this chapter, we present the experiments to validate our video summarization model and describe our dataset with our ground truth. Analysis of our results are also performed and are discussed independently. Validations of the video summarization are done manually while the trajectory clustering validation was performed by our ground truth.

The experiments were performed on an AMD Ryzen 7 1700 Eight-Core Processor 3.00 GHz with 16GB RAM, Windows 10 Pro 64 bits employing c++ language, our framework utilizes the following Python libraries: Scipy and Numpy (Walt *et al.*, 2011), Scikit-learn (Pedregosa *et al.*, 2011), PyWavelets (Lee G & Contributors, 2006). Some processes were parallelly executed (Graphic Card NVIDIA GeForce GTX 1070 Ti, in particular for a run the Self Organized Map using Tensorflow).

5.1 SSIG dataset

The SSIG-dataset was filmed in the Smart Sense Laboratory door. These videos are used for rare event detection. It contains only one view and it presents people in different situations, to name a few: People pointing in and pointing out of the laboratory, people closing and opening the door, people stopping for a long period and people walking outside of the laboratory, among other events. Figure 5.1 shows some frames extracted of our dataset and also different scenarios found on it. All videos have a resolution of 1280×720 colored frames. The ground truth for this dataset focuses on rare detection and finding the next list of activities defined as rare videos.

The criteria for defining normal videos are:

- People entering and leaving the laboratory by opening and closing the door.
- People opening the door for another people.

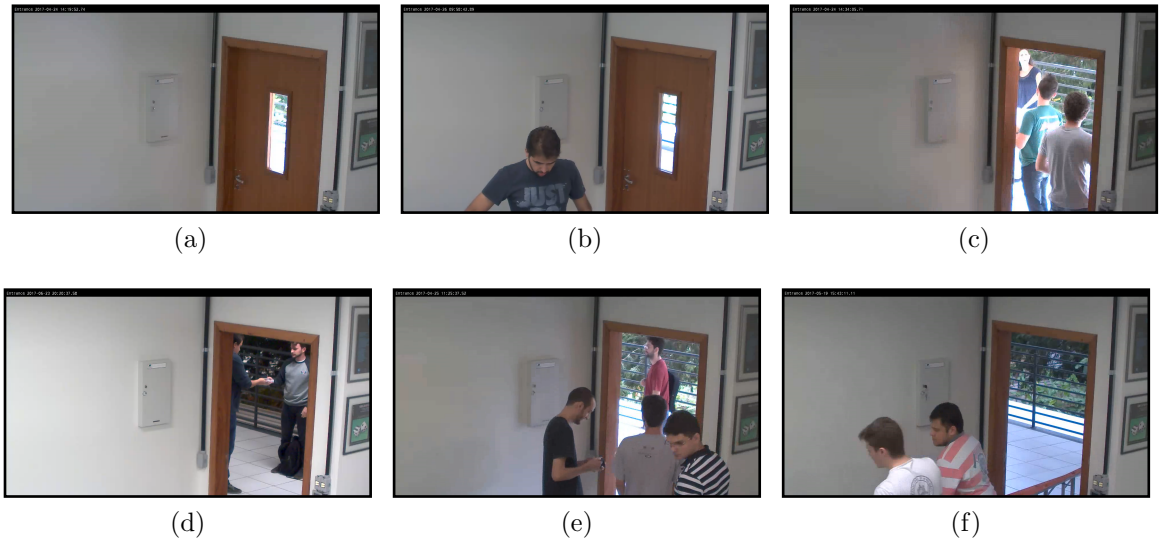
- A short amount of time spent in front of the camera (less than 10 seconds).
- People appearing quickly in some places where the camera can record.
- People starting the action of entering or leaving the laboratory, and in the middle of the action, changing direction or turning back.
- People going through the corridor outside the laboratory.
- People leaving and entering the side laboratory.

The criteria to define rare videos are the following:

- Using the key-box located near the door by long period.
- Standing in front of the camera for a long period (more than 10 seconds).
- Making several movements to come and repeatedly go to the laboratory.
- Leaning against the wall using a mobile devices (occurred several times in a single specific day).

From the present list of activities our approach focuses on the anomalies, it with the objective of showing rare events with priority, events as: (a) *The making of several movements to come and go to the laboratory*, (b) *stand in front of the camera for a long period of time or (c) using the key-box located near the door by long period of time.* It is important to highlight that a person should not leave from the focus of the camera since otherwise she or he will be considered as another person. For this ground truth, it is important that a person spends long periods of time in front of the camera to be considered rare.

The database consists of 5025 videos. The smallest video has a duration of two seconds while the largest has a duration of four minutes and twenty-three seconds. For the generation of this database, a static camera is recording the enter/exit door in the laboratory. Figure 5.1 depicts image examples of the camera view. Not all the videos of the dataset were used to experiment, without loss of generality we took 1000 videos to carry out our experiments. It should be noted that these videos were recorded continuously for two years. In each of the videos, it is possible to find anomalies as it is also possible not to find them. In general, these videos are of short duration, in most of them people appear. These videos were filmed at the door of the laboratory, in most of the videos, the upper part of the body of the people is visualized. The idea of this data set is to be able to segment a subset of this, that contains anomalous behaviors of a person in terms of their displacement in the video.



(a) (b) (c)

(d) (e) (f)

Figure 5.1: Frames extracted by SSIG-dataset.

5.1.1 Trajectory ground truth

Creating a ground truth is a tedious task and requires attention. One reason to create a dataset with its respective ground truth is that few study works found in literature for the study of trajectory provide the annotations. These extracted trajectories have a variable number of points. The number of points of each trajectory varies between six and 1500 points. Since no standard protocols were provided to generate the trajectory ground truth for explaining video summarization on this dataset, we adopted the following criteria: One set of trajectories is extracted to carry out the tests: Therefore, a set of a thousand trajectories from dataset were chosen. Also the ground truth trajectories was used to validate the clustering method based on their morphology. Furthermore, smooth morphology trajectories are separated from rare trajectories. As a result of this process, a set of 970 trajectories were obtained, this trajectories are clearly defined. The abnormal trajectories have abrupt changes in their morphology since they experienced changes of direction in small periods and low frequency in the dataset. The rare trajectories form a set of 30 trajectories, having in total 1000 of trajectories to be evaluated. This ground truth conforms our subset of data to perform our concept test to show the type of information that we can segment with our model.

5.1.2 Evaluation methodology

For the evaluation methodology, we propose two validation method, one for clustering and the other for video summarization. First we are going to describe the clustering validation method, we use this method to compare clustering methods. Then we explain how we perform the validation for our summarizer. These concepts are briefly described in the following paragraphs:

To compare the clustering methods used on our experiments, we use a method called *counting*. This method count how much elements of a cluster are wrong clustered. It to then extract a percentage of error for each cluster. With this information later we compute an average error of all clusters. To differentiate which elements in a cluster are wrong clustered, we take the preponderance of similar elements and define them as correct clustered trajectories, and all those differ from this biggest group will be defined as incorrect clustered elements. For example, Figure 5.2 shows an example of a generated cluster in which the wrong trajectories are boxed, it always considering its morphology. This cluster has a total of 22 elements, in which six of them are badly clustered, having a 27.27 % error percentage. This percentage is then averaged with the other cluster percentages in order to finally obtain a single percentage for each combination of clustering with descriptor methods.

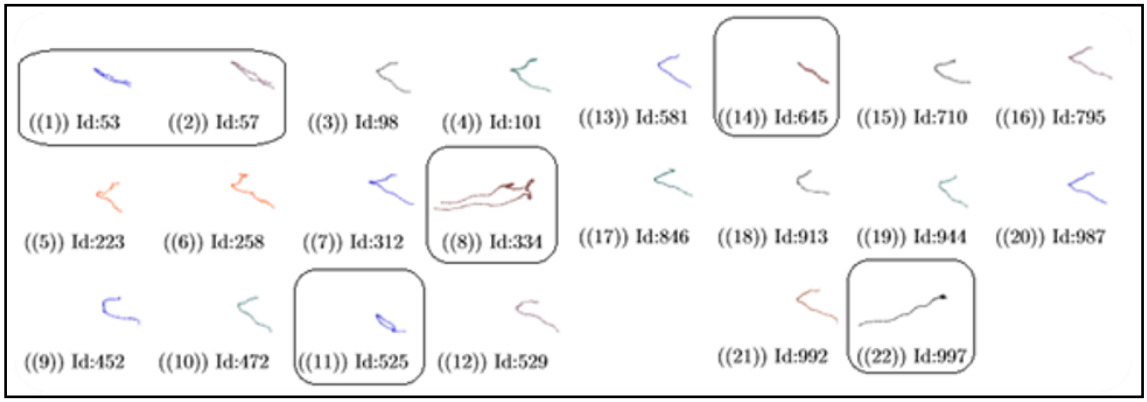


Figure 5.2: A cluster counting wrong clustered trajectories. It is an example of applying a counting method to a generated cluster, the wrong clustered trajectories are boxed.

For evaluate our summarization model, we compare our results with the proposed ground truth. For our ground truth as we mentioned before the trajectories were previously labeled as rare or normal. This information permits to evaluates what type of information is possible to segment with our approach. We define rare trajectories to the disorderly morphology behavior and the normal trajectories to the straightforward behavior. It means that normal trajectories can be easily described since these have relatively stable directions and predictable behaviors. Our evaluation analysis brings to identify false positives. In other words, during our experiments, some normal trajectories were identified as unusual. Our analysis also brings other measures such as the Specificity that allows us to evaluate the true negatives, which is usually passed out in most study works and it is relevant in the anomaly detection field.

According to Cunha (2011), there is no ideal method to validate the quality of generated video summaries. Each study creates its own validation methodology which often does not allow comparison with other scholarly article studies. The primary forms of validation of dynamic summaries are grouped into two categories: the objective metrics and validation through users. In agreement with Cunha (2011), one of the most latent problems in dynamic video summarization for surveillance is the lack of standardization of metrics, evaluation protocols, and datasets, which makes it difficult a comparison among methods. Some studies record their own datasets to perform

experiments, others use datasets from the literature but do not clarify how to reproduce the experiments.

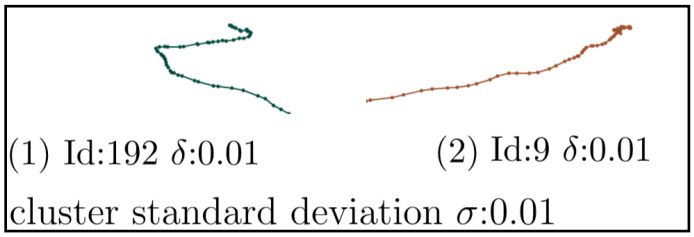
In addition to the quantitative measurements obtained, out study also provides qualitative measurements results. A tool used for this purpose included the t-Distributed Stochastic Neighbor Embedding (t-SNE) visualization method implemented in Sklearn python library.

5.1.3 Results

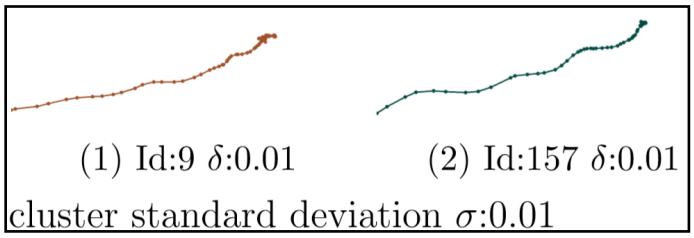
Now in this section, we explain our observations found in the evaluation step. Some of them are highlighting, the one that allowed us to improve our results for then to able to confirm our hypothesis.

The problem with interpolation is the approximation of one complicated trajectory by others that are more simple. These are the trajectories that are invariant, trajectories that do not have abrupt changes. The variability happens with the rare trajectories. They has pronounced changes in small intervals of time and interpolation cannot describe the behavior with sufficient accuracy. Figure 4.9 shows an example of this issue. The function obtained by the interpolation (red line) does not describe completely each real point (blue points) extracted from our experiment. The problem of inaccuracy in the polynomial interpolation was overcome with discrete Fourier and Wavelet transforms. Once the coefficients in each sub-trajectory are obtained, these coefficients can be used as characteristics of a vector after concatenating these two feature vectors. The characteristic vector is ready to be used on the clustering process. When the experiments were carried out with Fourier transform and Wavelet transform as descriptors, an improvement over interpolation was observed. As evidence of this experiment, the Figure 5.3 shows results obtained by applying Fourier transform and interpolation with the same dataset and using Kernel Density Estimation (KDE) as a grouper to pair similar trajectory features. This grouper matches one feature with the most similar feature from the universe of features provided in the input. For instance, the trajectory with *Id:9* shown in Figure 5.3 is grouping correctly with Fourier transform. This means that Fourier Transform as a descriptor extracts characteristics for grouping better than the interpolation method. This is then used as a description of the morphology of a trajectory.

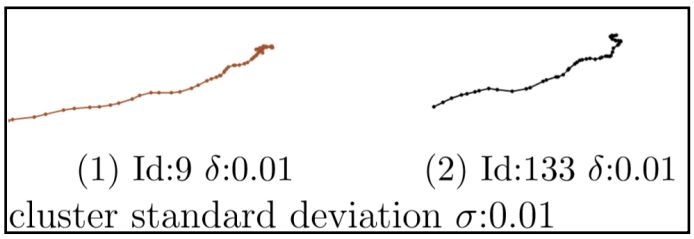
Due to an experiment with KDE, Fourier and Interpolation, the descriptors Fourier and Wavelet provide an improvement over interpolation. We can contrast this phenomenon in the Figure 5.3. Fourier and Wavelet descriptors provide similar results and to show the spaces generated by these two methods we decided to plot them by the same dataset. This experiment aims to discern the distribution of generated characteristics. For this task, we decided to use t-distributed Stochastic Neighbor Embedding. The t-SNE method reduced the dimensionality of our features from twenty-one to three dimensions. This worked only for visualization purposes of our descriptor performance. Figure 5.4b shows the space generated with the Fourier transform descriptor while Fig-



(a) Interpolation result.



(b) Fourier transform result.



(c) Wavelet transform result.

Figure 5.3: For our approach Fourier and Wavelet transforms extract robust features than interpolation.

ure 5.4c shows the space generated with the Wavelet transform descriptor. We conclude that the wavelet descriptor trends to distribute the data in a better way, producing a better separation of the characteristics extracted from each trajectory.

According to Table 5.1, the results show the percentage of error obtained with four combinations of descriptors with clustering. These results are obtained by applying the counting method. This provides a percentage error in each group so that finally at the end of the process, there are a percentage of trajectories that are incorrectly matched. For this experiment we used the wavelet and Fourier methods as a descriptor and SOM and affinity propagation as groupers. We can see that although the Affinity propagation algorithm generates a higher number of clusters, it appropriately groups the points on hyper-plane. Due to the measurements made with the counting method, we can observe that the Wavelet descriptor with the Affinity propagation as a grouper provides better results than the Self Organized Maps. Despite SOM being a good trajectory classifier, it has a defined number of clusters. This forces the group elements that are at the edges of the generated space to be included in a group, and does not create a new cluster like the Affinity propagation algorithm does.

We will explain the validation of our summarizer with the proposed ground truth. In order to evaluate our results, it is necessary to define a *threshold*, it marks the limit

	Clustering SOM (%)	Clustering Affinity (%)
Interpolation Descriptor	22.28	15.67
Fourier Descriptor	9.90	9.20
Wavelet Descriptor	8.77	**6.77**

Table 5.1: Error percentages found with the *Counting* method.

of the rare from the normal clusters. *The threshold is defined as the maximum number of elements that a cluster must have to be considered rare.* As we mentioned previously our ground truth consists of 970 normal and 30 rare trajectories to be evaluate. Table 5.2 shows our results obtained with different *thresholds*. It shows the results obtained by prioritizing the gathering of rare trajectories from the dataset.

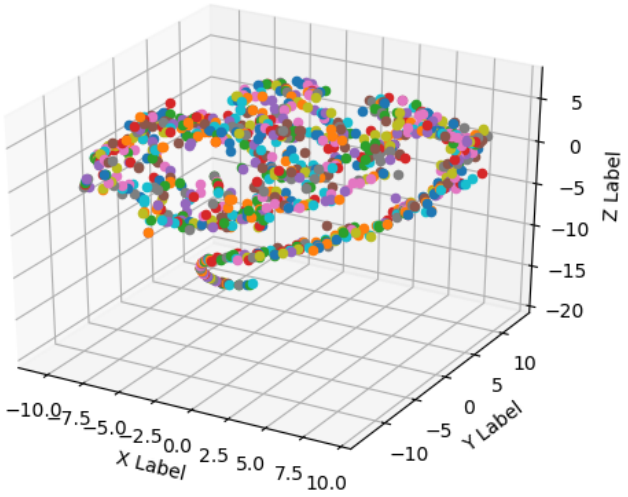
Table 5.2 shows that the *threshold* influences directly in the results, and if we choose an appropriate *threshold* for a determined set of data, our model obtains a maximum-accuracy. Thus it, depending on the *threshold* accuracy can change. If this value is moving away from the *threshold* which provide maximum-accuracy, the model detects normal trajectories as rare elements. On the other hand, if the *threshold* is still not near to the *threshold* which provide maximum-accuracy, this value is not enough to cover all rare trajectories. Part of the summarization consists of showing the relevant content and eliminating the redundant. Our approach segment the video information prioritizing the part that is considered relevant and rare, since rare videos form the summary of a videos in a surveillance domain.

Threshold	Accuracy	Precision	Recall	Specificity
1	0,991	1,000	0,700	1,000
2	**0,997**	**0,965**	**0,933**	**0,998**
3	0,989	0,731	1,000	0,988

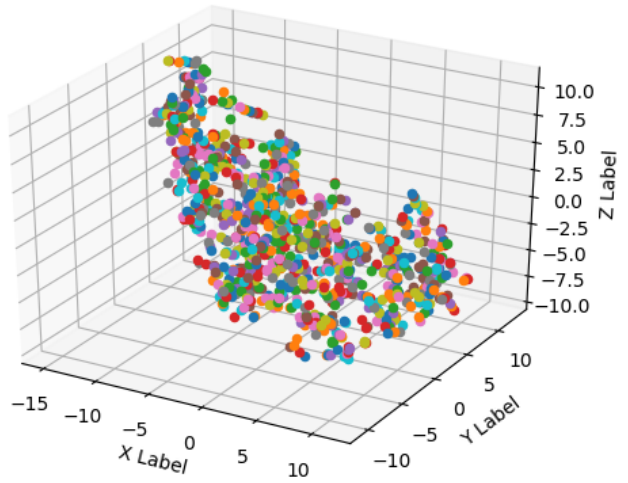
Table 5.2: Results of gathering rare trajectories by our model.

The error rate of false positives of a threshold with value three is 1.13 %. If we continue increasing the threshold value, this rate will increased considerably, the *Specificity* can express this behavior. However, our approach can prioritize rare videos.

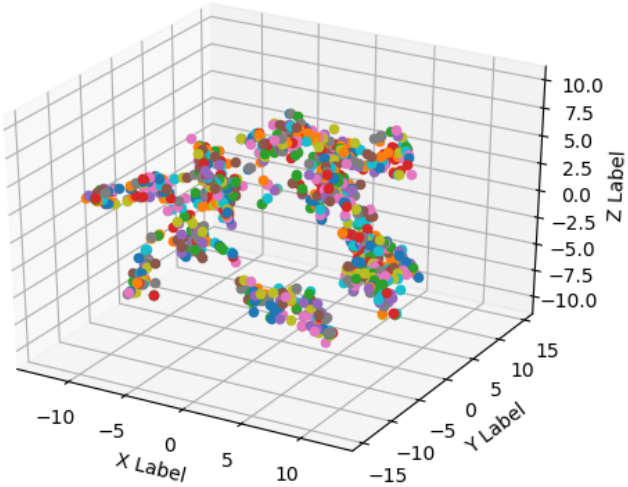
Finally, the labeling generated by our clustering method is presented on Figure 5.5a. This is plotted with t-SNE, the yellow points are considered rares. Figure 5.5b shows the rare trajectories detected by our model with threshold that considerate just one and two elements as limit and Figure 5.6 shows three clusters generated by our model, we can see that each cluster groups similar trajectories by morphology. The Figures 5.7 and 5.8 show results obtained after processing our model on surveillance videos. They report rare trajectories on their respective videos after gathering them. For visualization purposes a trajectory is written on the video. Each point of color represents a different position taken in *preliminaries* part. The white point represents the point taken for the generation of the trajectory. This point permits observation of the direction that the trajectory takes on each detected point.



(a) Interpolation descriptor result.

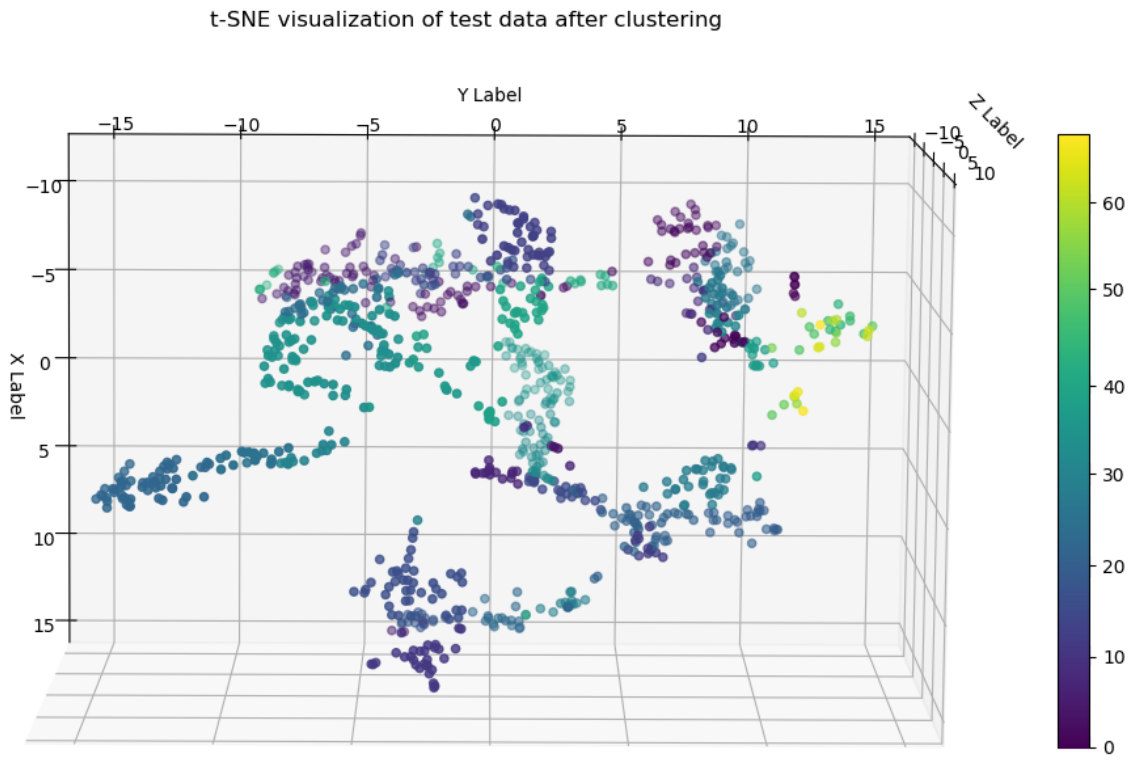


(b) Fourier transform descriptor result.

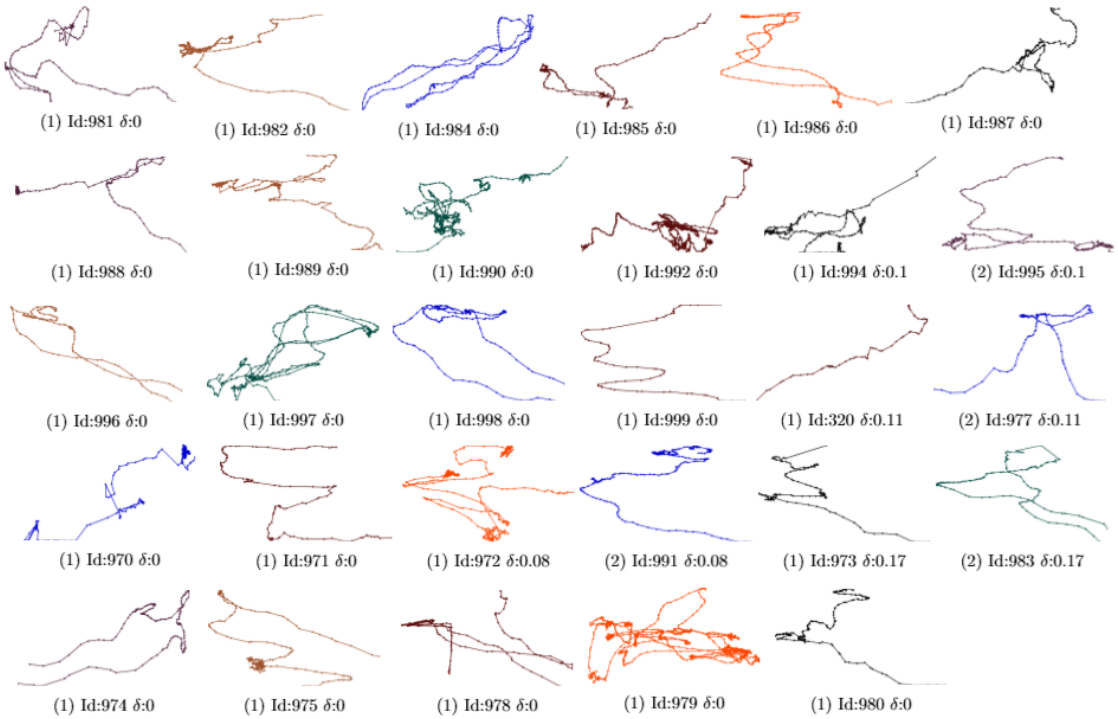


(c) Wavelet transform descriptor result.

Figure 5.4: Spaces generated by Polynomial Interpolation, DFT and DWT descriptors.



(a) Space generated after applying t-SNE, each color tonality represents a different cluster. Trajectories with yellow tonality are defined as rare by our model.



(b) Trajectories detected as rare by our model. The only trajectory that we did not define in our ground truth as rare, is the trajectory labeled with *Id:320*.

Figure 5.5: Final results by our model.

Figure 5.6: Normal trajectories detected by our model and three different generated clusters.

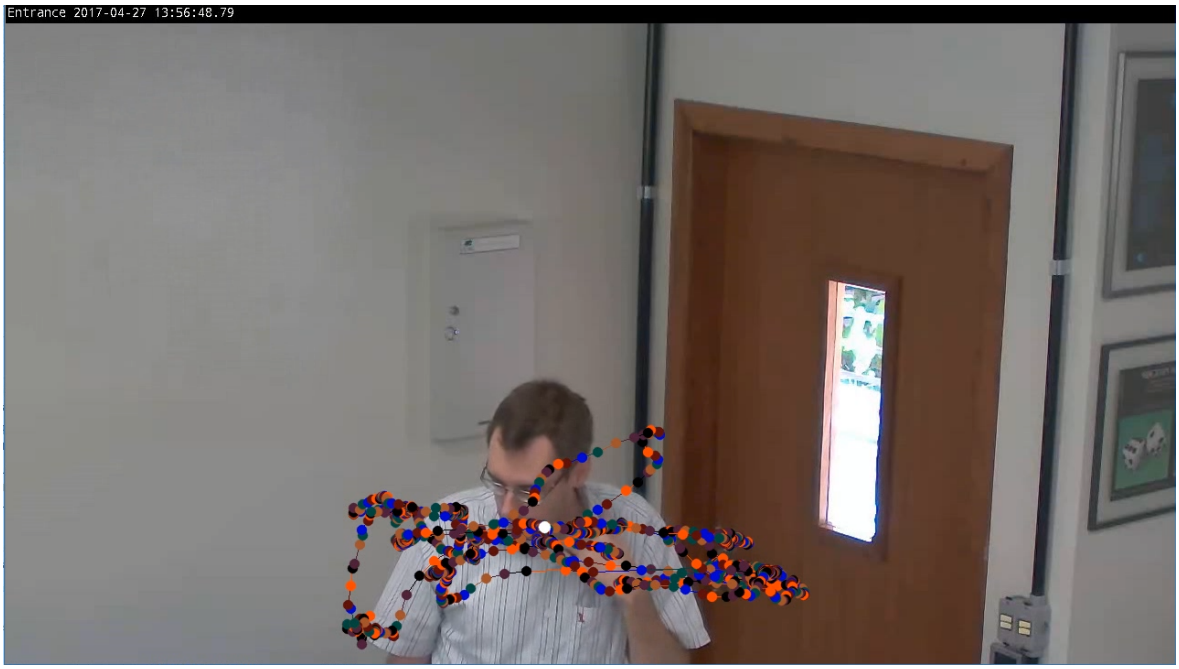


Figure 5.7: Thumbnail of a rare video with the rare trajectory number one. The white point represent the fiducial or reference point taken. This trajectory presents pronounced deformations. The trajectory was found and written on its corresponding video.

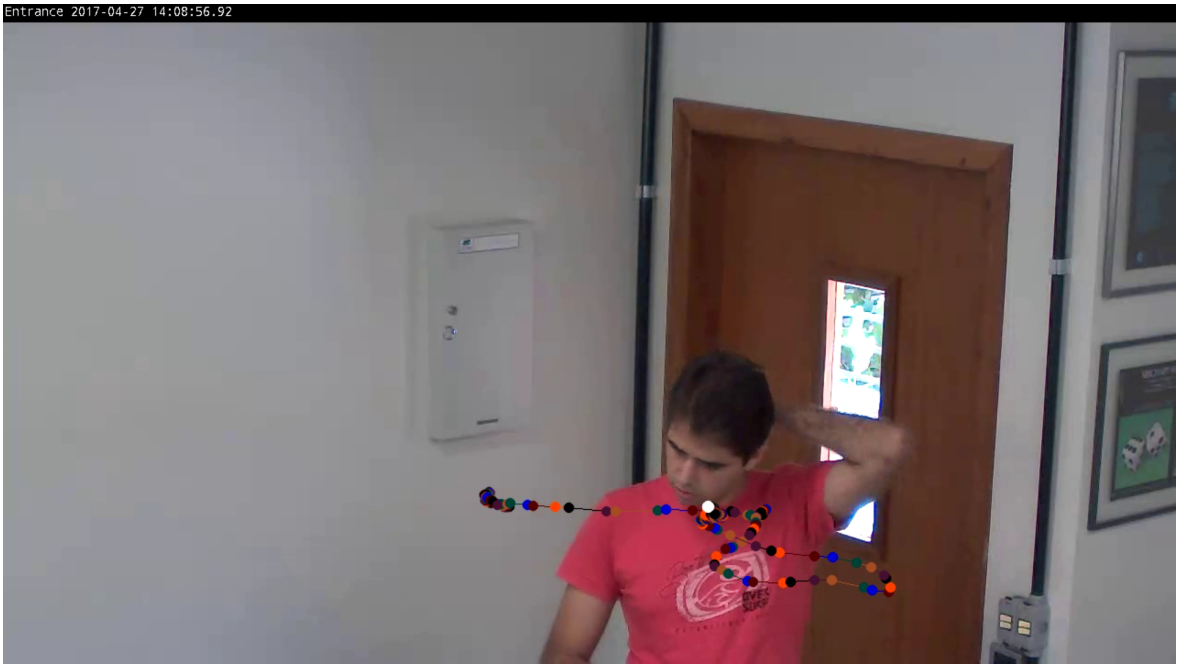


Figure 5.8: Thumbnail of a rare video with the rare trajectory number two, The white point represent the fiducial or reference point taken. The trajectory was found and written on its video.

Chapter 6

Conclusions and future works

This study presents a methodology for the generation of a dynamic summary of surveillance videos based on trajectories. For that purpose, we present a model which label video surveillance by human trajectory. It is with an emphasis on trajectory descriptors in conjunction with the unsupervised clustering method. Our feature vectors are based on real trajectories generated by people. This contributes to the literature found up until and is a unique model concerning the combination of techniques and objectives in summarization of surveillance video based on trajectories. In our opinion, it is correct to summarize based on rarity in the surveillance domain and to present the events according to the occurred frequency perceived on videos.

One of the most latent problems in dynamic video summarization for surveillance is the lack of standardization of metrics, evaluation protocols, and datasets, which makes it difficult a comparison among methods. While some study works record their own datasets to perform experiments, others use datasets from the literature but do not clarify how to reproduce the experiments. This study record own SSIG laboratory dataset to perform experiments and the *counting* method to compare clustering methods.

Given the difficulty of finding data-sets that work with video surveillance with trajectories, the proposal to use a new dataset arose. It is expected that the results are similar in other contexts; it means, in other places and people. The problems of gathering correct trajectories are treated by tracking and the pose-estimation methods that solve the problem considerably; for that, it is expected to not have additional problems on different environments as long as it is possible to visualize the majority of full body people. It is also appropriate to mention that the algorithms of people detection take longer time when the number of people is greater in the video.

The experimental results indicate the potential of this proposed method to summarize surveillance videos based on trajectories. The qualitative and quantitative analysis shows the validity of our model in a real video database. All this analysis is evidence that the trajectory information provides reliable video summarization information. This study proof that trajectories can add information to segment information

on videos.

Similar to the Fourier technique, we employ the Wavelet transform because of the variation in the signals that correspond to sub-trajectories o signal which describe the changes of a trajectory morphology. The t-Distributed Stochastic Neighbor Embedding method results appropriated for visualize our space of characteristics and even though the interpolation promises to adequately describe the sub-trajectories defined as functions, besides normalizing the number of points, The Wavelet transforms results as a better descriptor.

With the results of this study, we can notice that there are still some points of methodology to be explored. For example, the investigation of summarization of surveillance videos using cameras with movement. With each movement of the camera, new scenes are created and as a consequence, changes the meaning of the context. Other features to explore to add our model include visual features and motion features. We believe that it is still possible to improve our results and further our investigation into feature descriptors from trajectories and clustering methods.

As a final conclusion it can be affirmed that the aim was achieved as well as the sub-objective. We detected rare trajectories considering them as indivisible units, we investigated descriptors of trajectories based on morphology being wavelet descriptor with histograms our best option, we investigated unsupervised clustering techniques with unknown number of clusters to classify trajectories having better results with Affinity Propagation and finally the counting method was used to compare clustering methods and the experimentation with our ground truth.

References

Aggarwal, Jake K, & Ryoo, Michael S. 2011. Human activity analysis: A review. *ACM Computing Surveys (CSUR)*, **43**(3), 16.

Ajmal, Muhammad, Naseer, Mudasser, Ahmad, Farooq, & Saleem, Asma. 2017. Human motion trajectory analysis based video summarization. *Pages 550–555 of: 2017 16th IEEE International Conference on Machine Learning and Applications (ICMLA)*. IEEE.

Arthur, David, & Vassilvitskii, Sergei. 2007. k-means++: The advantages of careful seeding. *Pages 1027–1035 of: Proceedings of the eighteenth annual ACM-SIAM symposium on Discrete algorithms*. Society for Industrial and Applied Mathematics.

Austin, Anne, Barnard, Jonathan, & Hutcheon, Nicola. 2016. *Media Consumption Forecasts 2016*. Zenith The ROI Agency.

Bedagkar-Gala, Apurva, & Shah, Shishir K. 2014. A survey of approaches and trends in person re-identification. *Image and Vision Computing*, **32**(4), 270–286.

Bian, Jiang, Tian, Dayong, Tang, Yuanyan, & Tao, Dacheng. 2018. A survey on trajectory clustering analysis. *arXiv preprint arXiv:1802.06971*.

Cahuina, Edward Jorge Yuri Cayllahua. 2013. *A New Method for Static Video Summarization Using Visual Words and Video Temporal Segmentation*. Ms. dissertation, Universidad Federal de Ouro Preto.

Cao, Zhe, Simon, Tomas, Wei, Shih-En, & Sheikh, Yaser. 2017. Realtime Multi-Person 2D Pose Estimation using Part Affinity Fields. *In: Conference on Computer Vision and Pattern Recognition*.

Chabris, Christopher F, & Simons, Daniel. 2011. *The invisible gorilla: And other ways our intuitions deceive us*. Harmony.

Chaudhry, Rizwan, Ravichandran, Avinash, Hager, Gregory, & Vidal, René. 2009. Histograms of oriented optical flow and binet-cauchy kernels on nonlinear dynamical systems for the recognition of human actions. *Pages 1932–1939 of: Computer Vision and Pattern Recognition*. IEEE.

Chen, Yushi, Ma, Shunli, Chen, Xi, & Ghamisi, Pedram. 2017. Hyperspectral data clustering based on density analysis ensemble. *Remote Sensing Letters*, **8**(2), 194–203.

Cunha, Tiago Oliveira. 2011. *Sumarização automática de rushes vídeos baseada em características espaciais e espaço-temporais.* Ms. dissertation, Universidad Federal de Minas Gerais.

Dalal, Navneet, & Triggs, Bill. 2005. Histograms of oriented gradients for human detection. *Pages 886–893 of: Computer Vision and Pattern Recognition*, vol. 1. IEEE.

Damnjanovic, U., Fernandez, V., Izquierdo, E., & Martinez, J. M. 2008. Event Detection and Clustering for Surveillance Video Summarization Uros. *Ninth International Workshop on Image Analysis for Multimedia Interactive Services Event*, 63–66.

De Avila, Sandra Eliza Fontes, Lopes, Ana Paula Brandão, da Luz, Antonio, & de Albuquerque Araújo, Arnaldo. 2011. VSUMM: A mechanism designed to produce static video summaries and a novel evaluation method. *Pattern Recognition Letters*, **32**(1), 56–68.

Dee, Hannah M, & Velastin, Sergio A. 2008. How close are we to solving the problem of automated visual surveillance? *Machine Vision and Applications*, **19**.

Do, Chuong B, & Batzoglou, Serafim. 2008. What is the expectation maximization algorithm? *Nature biotechnology*, **26**(8), 897–899.

Eppler, Martin J, & Mengis, Jeanne. 2004. The concept of information overload: A review of literature from organization science, accounting, marketing, MIS, and related disciplines. *The information society*, **20**(5), 325–344.

Ergezer, Hamza, & Leblebicioğlu, Kemal. 2016. Anomaly detection and activity perception using covariance descriptor for trajectories. *Pages 728–742 of: European Conference on Computer Vision*. Springer.

Evangelopoulos, Georgios, Zlatintsi, Athanasia, Potamianos, Alexandros, Maragos, Petros, Rapantzikos, Konstantinos, Skoumas, Georgios, & Avrithis, Yannis. 2013. Multimodal saliency and fusion for movie summarization based on aural, visual, and textual attention. *IEEE Transactions on Multimedia*, **15**(7), 1553–1568.

Fortun, Denis, Bouthemy, Patrick, & Kervrann, Charles. 2015. Optical flow modeling and computation: a survey. *Computer Vision and Image Understanding*, **134**, 1–21.

Frey, Brendan J, & Dueck, Delbert. 2007. Clustering by passing messages between data points. *science*, **315**(5814), 972–976.

Furini, Marco, Geraci, Filippo, Montangero, Manuela, & Pellegrini, Marco. 2008. On Using Clustering Algorithms to Produce Video Abstracts for the Web Scenario. *IEEE Consumer Communications and Networking Conference*, 1112–1116.

Höferlin, M., Höferlin, B., Heidemann, G., & Weiskopf, D. 2013. Interactive Schematic Summaries for Faceted Exploration of Surveillance Video. *IEEE Transactions on multimedia*, **Vol. 15 nro. 04**, 908–920.

Hotelling, Harold. 1933. Analysis of a complex of statistical variables into principal components. *Journal of educational psychology*, **24**(6), 417.

Hu, Weiming, Li, Xi, Tian, Guodong, Maybank, Stephen, & Zhang, Zhongfei. 2013. An incremental DPMM-based method for trajectory clustering, modeling, and retrieval. *IEEE transactions on pattern analysis and machine intelligence*, **35**(5), 1051–1065.

IHS, Video Surveillance Research Area. 2016. Top video surveillance trends for 2016. *The IHS Technology*.

Jain, Anil K. 2010. Data clustering: 50 years beyond K-means. *Pattern recognition letters*, **31**(8), 651–666.

Ji, Zhong, Su, Yuting, Qian, Rongrong, & Ma, Jintao. 2010. Surveillance video summarization based on moving object detection and trajectory extraction. *International Conference on Signal Processing Systems*, **Vol 2**, 250–253.

Keval, Hina, & Sasse, Martina Angela. 2010. "Not the Usual Suspects": a study of factors reducing the effectiveness of CCTV. *Security Journal*, **23**(2), 134–154.

Khan, Yasmin S., & Pawar, Soudamini. 2015. Video Summarization: Survey on Event Detection and Summarization in Soccer Video. *International Journal of Advanced Computer Science and Applications*, **Vol. 06 nro. 11**, 256–259.

Kloss, Ricardo Barbosa, Cirne, Marcos Vinicius Mussel, Silva, Samira, Pedrini, Helio, & Schwartz, William Robson. 2015. Partial Least Squares Image Clustering. *SIBGRAPI - Conference on Graphics, Patterns and Images*.

Kohonen, Teuvo. 1998. The self-organizing map. *Neurocomputing*, **21**(1-3), 1–6.

Kong, Xiangjie, Li, Menglin, Ma, Kai, Tian, Kaiqi, Wang, Mengyuan, Ning, Zhaolong, & Xia, Feng. 2018. Big trajectory data: A survey of applications and services. *IEEE Access*, **6**, 58295–58306.

Lai, Po Kong, Décombas, Marc, Moutet, Kelvin, & Laganière, Robert. 2016. Video summarization of surveillance cameras. *Pages 286–294 of: Advanced Video and Signal Based Surveillance (AVSS), 2016 13th IEEE International Conference on*. IEEE.

Laxhammar, Rikard, & Falkman, Göran. 2014. Online learning and sequential anomaly detection in trajectories. *IEEE transactions on pattern analysis and machine intelligence*, **36**(6), 1158–1173.

Lee, Jae-Gil, Han, Jiawei, & Whang, Kyu-Young. 2007. Trajectory clustering: a partition-and-group framework. *Pages 593–604 of: Proceedings of the 2007 ACM SIGMOD international conference on Management of data*. ACM.

Lee G, Wasilewski F, Gommers R Wohlfahrt K O'Leary A Nahrstaedt H, & Contributors. 2006. *PyWavelets - Wavelet Transforms in Python*.

Li, Jing, & Allinson, Nigel M. 2008. A comprehensive review of current local features for computer vision. *Neurocomputing*, **71**(10), 1771–1787.

Liu, Bo, de Souza, Erico N, Matwin, Stan, & Sydow, Marcin. 2014. Knowledge-based clustering of ship trajectories using density-based approach. *Pages 603–608 of: Big Data (Big Data), 2014 IEEE International Conference on*. IEEE.

Liu, Xiaobai, Lin, Liang, Zhu, Song-Chun, & Jin, Hai. 2009. Trajectory parsing by cluster sampling in spatio-temporal graph. *Pages 739–746 of: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE.

Lowe, David G. 2004. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, **60**(2), 91–110.

Murugan, A Senthil, Devi, K Suganya, Sivaranjani, A, & Srinivasan, P. 2018. A study on various methods used for video summarization and moving object detection for video surveillance applications. *Multimedia Tools and Applications*, 1–18.

Naftel, Andrew, & Khalid, Shehzad. 2006. Motion trajectory learning in the DFT-coefficient feature space. *Pages 47–47 of: Computer Vision Systems, 2006 ICVS'06. IEEE International Conference on*. IEEE.

Pedregosa, Fabian, Varoquaux, Gaël, Gramfort, Alexandre, Michel, Vincent, Thirion, Bertrand, Grisel, Olivier, Blondel, Mathieu, Prettenhofer, Peter, Weiss, Ron, Dubourg, Vincent, *et al.* 2011. Scikit-learn: Machine learning in Python. *Journal of machine learning research*, **12**(Oct), 2825–2830.

Piciarelli, Claudio, Micheloni, Christian, & Foresti, Gian Luca. 2008. Trajectory-based anomalous event detection. *IEEE Transactions on Circuits and Systems for video Technology*, **18**(11), 1544–1554.

Pritch, Y., Ratovitch, S., Hendel, A., & Peleg, S. 2009. Clustered Synopsis of Surveillance Video. *Advanced Video and Signal Based Surveillance Clustered*, 195–200.

Rogers, Paul, Puryear, Rudy, & Root, James. 2013. Infobesity: The enemy of good decisions. *Bain Brief*, June.

Schreck, Tobias, Bernard, Jürgen, Von Landesberger, Tatiana, & Kohlhammer, Jörn. 2009. Visual cluster analysis of trajectory data with interactive kohonen maps. *Information Visualization*, **8**(1), 14–29.

Sebastian, Tinumol, & Puthiyidam, Jiby J. 2015. A survey on video summarization techniques. *Int. J. Comput. Appl*, **132**(13), 30–32.

Sillito, Rowland R, & Fisher, Robert B. 2008. Semi-supervised Learning for Anomalous Trajectory Detection. *Pages 035–1 of: BMVC*, vol. 1.

Sobral, Andrews, & Vacavant, Antoine. 2014. A comprehensive review of background subtraction algorithms evaluated with synthetic and real videos. *Computer Vision and Image Understanding*, **122**, 4–21.

Sodemann, Angela A, Ross, Matthew P, & Borghetti, Brett J. 2012. A review of anomaly detection in automated surveillance. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, **42**(6), 1257–1272.

Stutzer, Alois, & Zehnder, Michael. 2010. *Camera surveillance as a measure of counterterrorism?* Tech. rept. WWZ Discussion paper.

Tan, Hanlin, Zhai, Yongping, Liu, Yu, & Zhang, Maojun. 2016. Fast anomaly detection in traffic surveillance video based on robust sparse optical flow. *Pages 1976–1980 of: Acoustics, Speech and Signal Processing (ICASSP), 2016 IEEE International Conference on*. IEEE.

Tian, Dongkuan Xu. Yingjie. 2015. A Comprehensive Survey of Clustering Algorithms. *Annals of Data Science*.

Tint, Khin Thandar, & Soe, Kyi. 2013. Key frame extraction for video summarization using DWT wavelet statistics. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET)*, **2**(5), 1829–1833.

Téllez, Armando Martínez. 2009. *http://la-mecanica-cuantica.blogspot.com.br/2009/08/polinomios-de-legendre-aspectos.html*.

Torgerson, Warren S. 1952. Multidimensional scaling: I. Theory and method. *Psychometrika*, **17**(4), 401–419.

Torres, Berthin S, *et al.* 2016. Detection of complex video events based on visual rhythm. *The Visual Computer*, 1–22.

Truong, Ba Tu, & Venkatesh, Svetha. 2007. Video abstraction: A systematic review and classification. *ACM transactions on multimedia computing, communications, and applications (TOMM)*, **3**(1), 3.

Tung, Pham Thanh, & Ngoc, Ly Quoc. 2014. Elliptical density shape model for hand gesture recognition. 186–191.

Turchini, Francesco, Seidenari, Lorenzo, & Del Bimbo, Alberto. 2015. Understanding sport activities from correspondences of clustered trajectories. *Pages 43–50 of: Proceedings of the IEEE International Conference on Computer Vision Workshops*.

Vázquez-Martín, Ricardo, & Bandera, Antonio. 2013. Spatio-temporal feature-based keyframe detection from video shots using spectral clustering. *Pattern Recognition Letters*, **34**(7), 770–779.

Wallace, E., Diffley, C., Branch, Great Britain. Home Office. Police Scientific Development, & Home Office, St. Albans (GB). Police Scientific Development Branch. 1998. *CCTV: Making it work : CCTV control room rrgonomics*. Publication (Great Britain. Home Office. Police Scientific Development Branch). Police Scientific Development Branch.

Walt, Stéfan van der, Colbert, S Chris, & Varoquaux, Gael. 2011. The NumPy array: a structure for efficient numerical computation. *Computing in Science & Engineering*, **13**(2), 22–30.

Wang, Shizheng, Yang, Jianwei, Zhao, Yanyun, Cai, Anni, & Li, Stan Z. 2011. A surveillance video analysis and storage scheme for scalable synopsis browsing. *Pages 1947–1954 of: Computer Vision Workshops, 2011 IEEE International Conference on*. IEEE.

Welch, Greg, Bishop, Gary, *et al.* 1995. An introduction to the Kalman filter.

Welsh, BC, & Farrington, DP. 2008. *Effects of closed circuit television surveillance on crime. campbell Systematic Reviews*.

Weng, Shiuh-Ku, Kuo, Chung-Ming, & Tu, Shu-Kang. 2006. Video object tracking using adaptive Kalman filter. *Journal of Visual Communication and Image Representation*, **17**(6), 1190–1208.

Wu, Yi, Lim, Jongwoo, & Yang, Ming-Hsuan. 2013. Online object tracking: A benchmark. *Pages 2411–2418 of: Proceedings of the IEEE conference on computer vision and pattern recognition*.

Xu, Changsheng, Cheng, Jian, Zhang, Yi, Zhang, Yifan, & Lu, Hanqing. 2009. Sports Video Analysis: Semantics Extraction, Editorial Content Creation and Adaptation. *Journal of Multimedia*, **4**(2).

Xu, Hongteng, Zhou, Yang, Lin, Weiyao, & Zha, Hongyuan. 2015. Unsupervised trajectory clustering via adaptive multi-kernel-based shrinkage. *Pages 4328–4336 of: Proceedings of the IEEE International Conference on Computer Vision*.

Yilmaz, Alper, Javed, Omar, & Shah, Mubarak. 2006. Object tracking: A survey. *Acm computing surveys (CSUR)*, **38**(4), 13.

Zhang, Tianzhu, Lu, Hanqing, & Li, Stan Z. 2009. Learning semantic scene models by object classification and trajectory clustering. *Pages 1940–1947 of: Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*. IEEE.

Zhu, Yingying, Nayak, Nandita M., & Roy-Chowdhury, Amit K. 2013. Context-Aware Modeling and Recognition of Activities in Video. *Conference on Computer Vision and Pattern Recognition*, 2489–2496.